Proceedings Language Resources and Evaluation Conference (LREC) 2020

Nicoletta Calzolari, Frédéric Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph J Mariani, et al.

```
@Book{LREC:2020,
  editor    = {Nicoletta Calzolari  and  Frédéric Béchet  and
Philippe Blache  and  Khalid Choukri  and  Christopher Cieri  and
Thierry Declerck  and  Sara Goggi  and  Hitoshi Isahara  and  Bente
Maegaard  and  Joseph Mariani  and  Hélène Mazo  and  Asuncion
Moreno  and  Jan Odijk  and  Stelios Piperidis},
  title     = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1}
}

@InProceedings{yu-bohnet-poesio:2020:LREC,
  author    = {Yu, Juntao  and  Bohnet, Bernd  and  Poesio,
Massimo},
  title     = {Neural Mention Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1--10},
  abstract  = {Mention detection is an important preprocessing step
for annotation and interpretation in applications such as NER and
coreference resolution, but few stand-alone neural models have been
proposed able to handle the full range of mentions. In this work, we
propose and compare three neural network-based approaches to mention
detection. The first approach is based on the mention detection part
of a state of the art coreference resolution system; the second uses
ELMO embeddings together with a bidirectional LSTM and a biaffine
classifier; the third approach uses the recently introduced BERT
model. Our best model (using a biaffine classifier) achieves gains
of up to 1.8 percentage points on mention recall when compared with
a strong baseline in a HIGH RECALL coreference annotation setting.
The same model achieves improvements of up to 5.3 and 6.2 p.p. when
compared with the best-reported mention detection F1 on the CONLL
and CRAC coreference data sets respectively in a HIGH F1 annotation
setting. We then evaluate our models for coreference resolution by
using mentions predicted by our best model in start-of-the-art
coreference systems. The enhanced model achieved absolute
improvements of up to 1.7 and 0.7 p.p. when compared with our strong
baseline systems (pipeline system and end-to-end system)
respectively. For nested NER, the evaluation of our model on the
GENIA corpora shows that our model matches or outperforms state-of-
the-art models despite not being specifically designed for this
task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.1}
}

@InProceedings{wilkens-EtAl:2020:LREC,
```

```
  author    = {Wilkens, Rodrigo  and  Oberle, Bruno  and  Landragin,
Frédéric  and  Todirascu, Amalia},
  title     = {French Coreference for Spoken and Written Language},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {80--89},
  abstract  = {Coreference resolution aims at identifying and
grouping all mentions referring to the same entity. In French, most
systems run different setups, making their comparison difficult. In
this paper, we present an extensive comparison of several
coreference resolution systems for French. The systems have been
trained on two corpora (ANCOR for spoken language and Democrat for
written language) annotated with coreference chains, and augmented
with syntactic and semantic information. The models are compared
with different configurations (e.g. with and without singletons). In
addition, we evaluate mention detection and coreference resolution
apart. We present a full-stack model that outperforms other
approaches. This model allows us to study the impact of mention
detection errors on coreference resolution. Our analysis shows that
mention detection can be improved by focusing on boundary
identification while advances in the pronoun-noun relation detection
can help the coreference task. Another contribution of this work is
the first end-to-end neural French coreference resolution model
trained on Democrat (written texts), which compares to the state-of-
the-art systems for oral French.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.10}
}

@InProceedings{brunner-EtAl:2020:LREC,
  author    = {Brunner, Annelen  and  Engelberg, Stefan  and
Jannidis, Fotis  and  Tu, Ngoc Duyen Tanja  and  Weimer, Lukas},
  title     = {Corpus REDEWIEDERGABE},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {803--812},
  abstract  = {This article presents corpus REDEWIEDERGABE, a
German-language historical corpus with detailed annotations for
speech, thought and writing representation (ST\&WR). With
approximately 490,000 tokens, it is the largest resource of its
kind. It can be used to answer literary and linguistic research
questions and serve as training material for machine learning. This
paper describes the composition of the corpus and the annotation
structure, discusses some methodological decisions and gives basic
statistics about the forms of ST\&WR found in this corpus.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.100}
}
```

```
@InProceedings{egloff-picca:2020:LREC,
  author    = {Egloff, Mattia  and  Picca, Davide},
  title     = {WeDH - a Friendly Tool for Building Literary Corpora
Enriched with Encyclopedic Metadata},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {813--816},
  abstract  = {In recent years the interest in the use of
repositories of literary works has been successful. While many
efforts related to Linked Open Data go in the right direction, the
use of these repositories for the creation of text corpora enriched
with metadata remains difficult and cumbersome. In fact, many of
these repositories can be useful to the community not only for the
automatic creation of textual corpora but also for retrieving
crucial meta-information about texts. In particular, the use of
metadata provides the reader with a wealth of information that is
often not identifiable in the texts themselves. Our project aims to
fill both the access to the textual resources available on the web
and the possibility of combining these resources with sources of
metadata that can enrich the texts with useful information
lengthening the life and maintenance of the data itself. We
introduce here a user-friendly web interface of the Digital
Humanities toolkit named WeDH with which the user can leverage the
encyclopedic knowledge provided by DBpedia, wikidata and VIAF in
order to enrich the corpora with bibliographical and exegetical
knowledge. WeDH is a collaborative project and we invite anyone who
has ideas or suggestions regarding this procedure to reach out to
us.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.101}
}

@InProceedings{sabbatino-bostan-klinger:2020:LREC,
  author    = {Sabbatino, Valentino  and  Bostan, Laura Ana Maria
and  Klinger, Roman},
  title     = {Automatic Section Recognition in Obituaries},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {817--825},
  abstract  = {Obituaries contain information about people's values
across times and cultures, which makes them a useful resource for
exploring cultural history. They are typically structured similarly,
with sections corresponding to Personal Information, Biographical
Sketch, Characteristics, Family, Gratitude, Tribute, Funeral
Information and Other aspects of the person. To make this
information available for further studies, we propose a statistical
```

```
model which recognizes these sections. To achieve that, we collect a corpus of 20058 English obituaries from TheDaily Item, Remembering.CA and The London Free Press. The evaluation of our annotation guidelines with three annotators on 1008 obituaries shows a substantial agreement of Fleiss κ = 0.87. Formulated as an automatic segmentation task, a convolutional neural network outperforms bag-of-words and embedding-based BiLSTMs and BiLSTM-CRFs with a micro F1 = 0.81.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.102}
}

@InProceedings{stymne-stman:2020:LREC,
  author    = {Stymne, Sara  and  Östman, Carin},
  title     = {SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {826--834},
  abstract  = {We describe a new corpus, SLäNDa, the Swedish Literary corpus of Narrative and Dialogue. It contains Swedish literary fiction, which has been manually annotated for cited materials, with a focus on dialogue. The annotation covers excerpts from eight Swedish novels written between 1879--1940, a period of modernization of the Swedish language. SLäNDa contains annotations for all cited materials that are separate from the main narrative, like quotations and signs. The main focus is on dialogue, for which we annotate speech segments, speech tags, and speakers. In this paper we describe the annotation protocol and procedure and show that we can reach a high inter-annotator agreement. In total, SLäNDa contains annotations of 44 chapters with over 220K tokens. The annotation identified 4,733 instances of cited material and 1,143 named speaker--speech mappings. The corpus is useful for developing computational tools for different types of analysis of literary narrative and speech. We perform a small pilot study where we show how our annotation can help in analyzing language change in Swedish. We find that a number of common function words have their modern version appear earlier in speech than in narrative.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.103}
}

@InProceedings{papay-pad:2020:LREC,
  author    = {Papay, Sean  and  Padó, Sebastian},
  title     = {RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {835--841},
  abstract  = {We introduce RiQuA (RIch QUotation Annotations), a
corpus that provides quotations, including their interpersonal
structure (speakers and addressees) for English literary text. The
corpus comprises 11 works of 19th-century literature that were
manually doubly annotated for direct and indirect quotations. For
each quotation, its span, speaker, addressee, and cue are identified
(if present). This provides a rich view of dialogue structures not
available from other available corpora. We detail the process of
creating this dataset, discuss the annotation guidelines, and
analyze the resulting corpus in terms of inter-annotator agreement
and its properties. RiQuA, along with its annotations guidelines and
associated scripts, are publicly available for use, modification,
and experimentation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.104}
}

@InProceedings{schneider:2020:LREC,
  author    = {Schneider, Roman},
  title     = {A Corpus Linguistic Perspective on Contemporary
German Pop Lyrics with the Multi-Layer Annotated "Songkorpus"},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {842--848},
  abstract  = {Song lyrics can be considered as a text genre that
has features of both written and spoken discourse, and potentially
provides extensive linguistic and cultural information to scientists
from various disciplines. However, pop songs play a rather
subordinate role in empirical language research so far - most likely
due to the absence of scientifically valid and sustainable
resources. The present paper introduces a multiply annotated corpus
of German lyrics as a publicly available basis for multidisciplinary
research. The resource contains three types of data for the
investigation and evaluation of quite distinct phenomena: TEI-
compliant song lyrics as primary data, linguistically and literary
motivated annotations, and extralinguistic metadata. It promotes
empirically/statistically grounded analyses of genre-specific
features, systemic-structural correlations and tendencies in the
texts of contemporary pop music. The corpus has been stratified into
thematic and author-specific archives; the paper presents some basic
descriptive statistics, as well as the public online frontend with
its built-in evaluation forms and live visualisations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.105}
}

@InProceedings{grilo-EtAl:2020:LREC,
  author    = {Grilo, Sara  and  Bolrinha, Márcia  and  Silva, João
and  Vaz, Rui  and  Branco, António},
  title     = {The BDCamões Collection of Portuguese Literary
Documents: a Research Resource for Digital Humanities and Language
```

Technology},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {849--854},
  abstract = {This paper presents the BDCamões Collection of Portuguese Literary Documents, a new corpus of literary texts written in Portuguese that in its inaugural version includes close to 4 million words from over 200 complete documents from 83 authors in 14 genres, covering a time span from the 16th to the 21st century, and adhering to different orthographic conventions. Many of the texts in the corpus have also been automatically parsed with state-of-the-art language processing tools, forming the BDCamões Treebank subcorpus. This set of characteristics makes of BDCamões an invaluable resource for research in language technology (e.g. authorship detection, genre classification, etc.) and in language science and digital humanities (e.g. comparative literature, diachronic linguistics, etc.).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.106}
}

@InProceedings{frossard-EtAl:2020:LREC,
  author   = {Frossard, Esteban and Coustaty, Mickael and Doucet, Antoine and Jatowt, Adam and Hengchen, Simon},
  title   = {Dataset for Temporal Analysis of English-French Cognates},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {855--859},
  abstract = {Languages change over time and, thanks to the abundance of digital corpora, their evolutionary analysis using computational techniques has recently gained much research attention. In this paper, we focus on creating a dataset to support investigating the similarity in evolution between different languages. We look in particular into the similarities and differences between the use of corresponding words across time in English and French, two languages from different linguistic families yet with shared syntax and close contact. For this we select a set of cognates in both languages and study their frequency changes and correlations over time. We propose a new dataset for computational approaches of synchronized diachronic investigation of language pairs, and subsequently show novel findings stemming from the cognate-focused diachronic comparison of the two chosen languages. To the best of our knowledge, the present study is the first in the literature to use computational approaches and large data to make a cross-language diachronic analysis.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.107}

}

@InProceedings{waldisphl-dannells-borin:2020:LREC,
  author   = {Waldispühl, Michelle and Dannells, Dana and Borin, Lars},
  title   = {Material Philology Meets Digital Onomastic Lexicography: The NordiCon Database of Medieval Nordic Personal Names in Continental Sources},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {860--867},
  abstract = {We present NordiCon, a database containing medieval Nordic personal names attested in Continental sources. The database combines formally interpreted and richly interlinked onomastic data with digitized versions of the medieval manuscripts from which the data originate and information on the tokens' context. The structure of NordiCon is inspired by other online historical given name dictionaries. It takes up challenges reported on in previous works, such as how to cover material properties of a name token and how to define lemmatization principles, and elaborates on possible solutions. The lemmatization principles for NordiCon are further developed in order to facilitate the connection to other name dictionaries and corpuses, and the integration of the database into Språkbanken Text, an infrastructure containing modern and historical written data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.108}
}

@InProceedings{mohammad:2020:LREC1,
  author   = {Mohammad, Saif M.},
  title   = {NLP Scholar: A Dataset for Examining the State of NLP Research},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {868--877},
  abstract = {Google Scholar is the largest web search engine for academic literature that also provides access to rich metadata associated with the papers. The ACL Anthology (AA) is the largest repository of articles on Natural Language Processing (NLP). We extracted information from AA for about 44 thousand NLP papers and identified authors who published at least three papers there. We then extracted citation information from Google Scholar for all their papers (not just their AA papers). This resulted in a dataset of 1.1 million papers and associated Google Scholar information. We aligned the information in the AA and Google Scholar datasets to create the NLP Scholar Dataset -- a single unified source of

```
information (from both AA and Google Scholar) for tens of thousands
of NLP papers. It can be used to identify broad trends in
productivity, focus, and impact of NLP research. We present here
initial work on analyzing the volume of research in NLP over the
years and identifying the most cited papers in NLP. We also list a
number of additional potential applications.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.109}
}

@InProceedings{aloraini-poesio:2020:LREC,
  author    = {Aloraini, Abdulrahman  and  Poesio, Massimo},
  title     = {Cross-lingual Zero Pronoun Resolution},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {90--98},
  abstract  = {In languages like Arabic, Chinese, Italian, Japanese,
Korean, Portuguese, Spanish, and many others, predicate arguments in
certain syntactic positions are not realized instead of being
realized as overt pronouns, and are thus called zero- or null-
pronouns. Identifying and resolving such omitted arguments is
crucial to machine translation, information extraction and other NLP
tasks, but depends heavily on semantic coherence and lexical
relationships. We propose a BERT-based cross-lingual model for zero
pronoun resolution, and evaluate it on the Arabic and Chinese
portions of OntoNotes 5.0. As far as we know, ours is the first
neural model of zero-pronoun resolution for Arabic; and our model
also outperforms the state-of-the-art for Chinese. In the paper we
also evaluate BERT feature extraction and fine-tune models on the
task, and compare them with our model. We also report on an
investigation of BERT layers indicating which layer encodes the most
suitable representation for the task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.11}
}

@InProceedings{virk-EtAl:2020:LREC,
  author    = {Virk, Shafqat Mumtaz  and  Hammarström, Harald  and
Forsberg, Markus  and  Wichmann, Søren},
  title     = {The DReaM Corpus: A Multilingual Annotated Corpus of
Grammars for the World's Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {878--884},
  abstract  = {There exist as many as 7000 natural languages in the
world, and a huge number of documents describing those languages
have been produced over the years. Most of those documents are in
paper format. Any attempts to use modern computational techniques
```

and tools to process those documents will require them to be digitized first. In this paper, we report a multilingual digitized version of thousands of such documents searchable through some well-established corpus infrastructures. The corpus is annotated with various meta, word, and text level attributes to make searching and analysis easier and more useful.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.110}
}

@InProceedings{mller-tikhonov-meyer:2020:LREC,
  author   = {Müller, Klaus and Tikhonov, Aleksej and Meyer, Roland},
  title   = {LiViTo: Linguistic and Visual Features Tool for Assisted Analysis of Historic Manuscripts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {885--890},
  abstract = {We propose a mixed methods approach to the identification of scribes and authors in handwritten documents, and present LiViTo, a software tool which combines linguistic insights and computer vision techniques in order to assist researchers in the analysis of handwritten historical documents. Our research shows that it is feasible to train neural networks for the automatic transcription of handwritten documents and to use these transcriptions as input for further learning processes. Hypotheses about scribes can be tested effectively by extracting visual handwriting features and clustering them appropriately. Methods from linguistics and from computer vision research integrate into a mixed methods system, with benefits on both sides. LiViTo was trained with historical Czech texts by 18th century immigrants to Berlin, a total of 564 pages from a corpus of about 5000 handwritten pages without indication of author or scribe. We provide an overview of the three-year development of LiViTo and an introduction into its methodology and its functions. We then present our findings concerning the corpus of Berlin Czech manuscripts and discuss possible further usage scenarios.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.111}
}

@InProceedings{abrami-stoeckel-mehler:2020:LREC,
  author   = {Abrami, Giuseppe and Stoeckel, Manuel and Mehler, Alexander},
  title   = {TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {891--900},
  abstract  = {The annotation of texts and other material in the
field of digital humanities and Natural Language Processing (NLP) is
a common task of research projects. At the same time, the annotation
of corpora is certainly the most time- and cost-intensive component
in research projects and often requires a high level of expertise
according to the research interest. However, for the annotation of
texts, a wide range of tools is available, both for automatic and
manual annotation. Since the automatic pre-processing methods are
not error-free and there is an increasing demand for the generation
of training data, also with regard to machine learning, suitable
annotation tools are required. This paper defines criteria of
flexibility and efficiency of complex annotations for the assessment
of existing annotation tools. To extend this list of tools, the
paper describes TextAnnotator, a browser-based, multi-annotation
system, which has been developed to perform platform-independent
multimodal annotations and annotate complex textual structures. The
paper illustrates the current state of development of TextAnnotator
and demonstrates its ability to evaluate annotation quality (inter-
annotator agreement) at runtime. In addition, it will be shown how
annotations of different users can be performed simultaneously and
collaboratively on the same document from different platforms using
UIMA as the basis for annotation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.112}
}

@InProceedings{gyawali-anastasiou-knoth:2020:LREC,
  author    = {Gyawali, Bikash  and  Anastasiou, Lucas  and  Knoth,
Petr},
  title     = {Deduplication of Scholarly Documents using Locality
Sensitive Hashing and Word Embeddings},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {901--910},
  abstract  = {Deduplication is the task of identifying near and
exact duplicate data items in a collection. In this paper, we
present a novel method for deduplication of scholarly documents. We
develop a hybrid model which uses structural similarity (locality
sensitive hashing) and meaning representation (word embeddings) of
document texts to determine (near) duplicates. Our collection
constitutes a subset of multidisciplinary scholarly documents
aggregated from research repositories. We identify several issues
causing data inaccuracies in such collections and motivate the need
for deduplication. In lack of existing dataset suitable for study of
deduplication of scholarly documents, we create a ground truth
dataset of $100K$ scholarly documents and conduct a series of
experiments to empirically establish optimal values for the
parameters of our deduplication method. Experimental evaluation
shows that our method achieves a macro F1-score of 0.90. We
productionise our method as a publicly accessible web API service
```

```
serving deduplication of scholarly documents in real time.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.113}
}

@InProceedings{boschetti-EtAl:2020:LREC,
  author    = {Boschetti, Federico  and  de felice, irene  and  Dei
Rossi, Stefano  and  Dell'Orletta, Felice  and  Di Giorgio, Michele
and  Miliani, Martina  and  Passaro, Lucia C.  and  Puddu, Angelica
and  Venturi, Giulia  and  Labanca, Nicola  and  Lenci, Alessandro
and  Montemagni, Simonetta},
  title     = {“Voices of the Great War”: A Richly Annotated Corpus
of Italian Texts on the First World War},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {911--918},
  abstract  = {“Voices of the Great War” is the first large corpus
of Italian historical texts dating back to the period of First World
War. This corpus differs from other existing resources in several
respects. First, from the linguistic point of view it gives account
of the wide range of varieties in which Italian was articulated in
that period, namely from a diastratic (educated vs. uneducated
writers), diaphasic (low/informal vs. high/formal registers) and
diatopic (regional varieties, dialects) points of view. From the
historical perspective, through a collection of texts belonging to
different genres it represents different views on the war and the
various styles of narrating war events and experiences. The final
corpus is balanced along various dimensions, corresponding to the
textual genre, the language variety used, the author type and the
typology of conveyed contents. The corpus is fully annotated with
lemmas, part-of-speech, terminology, and named entities. Significant
corpus samples representative of the different “voices” have also
been enriched with meta-linguistic and syntactic information. The
layer of syntactic annotation forms the first nucleus of an Italian
historical treebank complying with the Universal Dependencies
standard. The paper illustrates the final resource, the methodology
and tools used to build it, and the Web Interface for navigating
it.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.114}
}

@InProceedings{lapesa-EtAl:2020:LREC,
  author    = {Lapesa, Gabriella  and  Blessing, Andre  and
Blokker, Nico  and  Dayanik, Erenay  and  Haunss, Sebastian  and
Kuhn, Jonas  and  Padó, Sebastian},
  title     = {DEbateNet-mig15:Tracing the 2015 Immigration Debate
in Germany Over Time},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {919--927},
  abstract  = {DEbateNet-migr15 is a manually annotated dataset for
German which covers the public debate on immigration in 2015. The
building block of our annotation is the political science notion of
a claim, i.e., a statement made by a political actor (a politician,
a party, or a group of citizens) that a specific action should be
taken (e.g., vacant flats should be assigned to refugees). We
identify claims in newspaper articles, assign them to actors and
fine-grained categories and annotate their polarity and date. The
aim of this paper is two-fold: first, we release the full DEbateNet-
mig15 corpus and document it by means of a quantitative and
qualitative analysis; second, we demonstrate its application in a
discourse network analysis framework, which enables us to capture
the temporal dynamics of the political debate},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.115}
}

@InProceedings{lvarezmellado:2020:LREC,
  author    = {Álvarez-Mellado, Elena},
  title     = {A Corpus of Spanish Political Speeches from 1937 to
2019},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {928--932},
  abstract  = {This paper documents a corpus of political speeches
in Spanish. The documents in the corpus belong to the Christmas
speeches that have been delivered yearly by the head of state of
Spain since 1937. The historical period covered by these speeches
ranges from the Spanish Civil War and the Francoist dictatorship up
until today. As a result, the corpus reflects some of the most
significant events and political changes in the recent history of
Spain. Up until now, the speeches as a whole had not been collected
into a single, systematic and reusable resource, as most of the
texts were scattered among different sources. The paper describes:
(1) the composition of the corpus; (2) the Python interface that
facilitates querying and analyzing the corpus using the NLTK and
spaCy libraries and (3) a set of HTML visualizations aimed at the
general public to navigate the corpus and explore differences
between TF-IDF frequencies.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.116}
}

@InProceedings{cecchini-korkiakangas-passarotti:2020:LREC,
  author    = {Cecchini, Flavio Massimiliano  and  Korkiakangas,
Timo  and  Passarotti, Marco},
  title     = {A New Latin Treebank for Universal Dependencies:
Charters between Ancient Latin and Romance Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
```

Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {933--942},
  abstract = {The present work introduces a new Latin treebank that follows the Universal Dependencies (UD) annotation standard. The treebank is obtained from the automated conversion of the Late Latin Charter Treebank 2 (LLCT2), originally in the Prague Dependency Treebank (PDT) style. As this treebank consists of Early Medieval legal documents, its language variety differs considerably from both the Classical and Medieval learned varieties prevalent in the other currently available UD Latin treebanks. Consequently, besides significant phenomena from the perspective of diachronic linguistics, this treebank also poses several challenging technical issues for the current and future syntactic annotation of Latin in the UD framework. Some of the most relevant cases are discussed in depth, with comparisons between the original PDT and the resulting UD annotations. Additionally, an overview of the UD-style structure of the treebank is given, and some diachronic aspects of the transition from Latin to Romance languages are highlighted.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.117}
}

@InProceedings{rochasouza-EtAl:2020:LREC,
  author   = {Rocha Souza, Renato and Dorn, Amelie and Piringer, Barbara and Wandl-Vogt, Eveline},
  title   = {Identification of Indigenous Knowledge Concepts through Semantic Networks, Spelling Tools and Word Embeddings},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {943--947},
  abstract = {In order to access indigenous, regional knowledge contained in language corpora, semantic tools and network methods are most typically employed. In this paper we present an approach for the identification of dialectal variations of words, or words that do not pertain to High German, on the example of non-standard language legacy collection questionnaires of the Bavarian Dialects in Austria (DBÖ). Based on selected cultural categories relevant to the wider project context, common words from each of these cultural categories and their lemmas using GermaLemma were identified. Through word embedding models the semantic vicinity of each word was explored, followed by the use of German Wordnet (Germanet) and the Hunspell tool. Whilst none of these tools have a comprehensive coverage of standard German words, they serve as an indication of dialects in specific semantic hierarchies. Methods and tools applied in this study may serve as an example for other similar projects dealing with non-standard or endangered language collections, aiming to access, analyze and ultimately preserve native regional language

```
heritage.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.118}
}

@InProceedings{saleva:2020:LREC,
  author    = {Saleva, Jonne},
  title     = {A Multi-Orthography Parallel Corpus of Yiddish
Nouns},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {948--952},
  abstract  = {Yiddish is a low-resource language belonging to the
Germanic language family and written using the Hebrew alphabet. As a
language, Yiddish can be considered resource-poor as it lacks both
public accessible corpora and a widely-used standard orthography,
with various countries and organizations influencing the spellings
speakers use. While existing corpora of Yiddish text do exist, they
are often only written in a single, potentially non-standard
orthography, with no parallel version with standard orthography
available. In this work, we introduce the first multi-orthography
parallel corpus of Yiddish nouns built by scraping word entries from
Wiktionary. We also demonstrate how the corpus can be used to
bootstrap a transliteration model using the Sequitur-G2P grapheme-
to-phoneme conversion toolkit to map between various orthographies.
Our trained system achieves error rates between 16.79\% and 28.47\%
on the test set, depending on the orthographies considered. In
addition to quantitative analysis, we also conduct qualitative error
analysis of the trained system, concluding that non-phonetically
spelled Hebrew words are the largest cause of error. We conclude
with remarks regarding future work and release the corpus and
associated code under a permissive license for the larger community
to use.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.119}
}

@InProceedings{loiciga-hardmeier-sayeed:2020:LREC,
  author    = {Loáiciga, Sharid  and  Hardmeier, Christian  and
Sayeed, Asad},
  title     = {Exploiting Cross-Lingual Hints to Discover Event
Pronouns},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {99--103},
  abstract  = {Non-nominal co-reference is much less studied than
nominal coreference, partly because of the lack of annotated
corpora. We explore the possibility to exploit parallel multilingual
```

```
corpora as a means of cheap supervision for the classification of
three different readings of the English pronoun 'it': entity, event
or pleonastic, from their translation in several languages. We found
that the 'event' reading is not very frequent, but can be easily
predicted provided that the construction used to translate the 'it'
example is a pronoun as well. These cases, nevertheless, are not
enough to generalize to other types of non-nominal reference.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.12}
}

@InProceedings{gerhalter-EtAl:2020:LREC,
  author    = {Gerhalter, Katharina  and  Schneider, Gerlinde  and
Pollin, Christopher  and  Hummel, Martin},
  title     = {An Annotated Corpus of Adjective-Adverb Interfaces in
Romance Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {953--957},
  abstract  = {The final outcome of the project Open Access
Database: Adjective-Adverb Interfaces in Romance is an annotated and
lemmatised corpus of various linguistic phenomena related to Romance
adjectives with adverbial functions. The data is published under
open-access and aims to serve linguistic research based on
transparent and accessible corpus-based data. The annotation model
was developed to offer a cross-linguistic categorization model for
the heterogeneous word-class “adverb”, based on its diverse forms,
functions and meanings. The project focuses on the interoperability
and accessibility of data, with particular respect to reusability in
the sense of the FAIR Data Principles. Topics presented by this
paper include data compilation and creation, annotation in XML/TEI,
data preservation and publication process by means of the GAMS
repository and accessibility via a search interface. These aspects
are tied together by semantic technologies, using an ontology-based
approach.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.120}
}

@InProceedings{ehrmann-EtAl:2020:LREC,
  author    = {Ehrmann, Maud  and  Romanello, Matteo  and
Clematide, Simon  and  Ströbel, Phillip Benjamin  and  Barman,
Raphaël},
  title     = {Language Resources for Historical Newspapers: the
Impresso Collection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {958--968},
```

```
  abstract  = {Following decades of massive digitization, an
unprecedented amount of historical document facsimiles can now be
retrieved and accessed via cultural heritage online portals. If this
represents a huge step forward in terms of preservation and
accessibility, the next fundamental challenge-- and real promise of
digitization-- is to exploit the contents of these digital assets,
and therefore to adapt and develop appropriate language technologies
to search and retrieve information from this `Big Data of the Past'.
Yet, the application of text processing tools on historical
documents in general, and historical newspapers in particular, poses
new challenges, and crucially requires appropriate language
resources. In this context, this paper presents a collection of
historical newspaper data sets composed of text and image resources,
curated and published within the context of the `impresso - Media
Monitoring of the Past' project. With corpora, benchmarks, semantic
annotations and language models in French, German and Luxembourgish
covering ca. 200 years, the objective of the impresso resource
collection is to contribute to historical language resources, and
thereby strengthen the robustness of approaches to non-standard
inputs and foster efficient processing of historical documents.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.121}
}

@InProceedings{kampe-duan-hahn:2020:LREC,
  author    = {Kampe, Bernd  and  Duan, Tinghui  and  Hahn, Udo},
  title     = {Allgemeine Musikalische Zeitung as a Searchable
Online Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {969--976},
  abstract  = {The massive digitization efforts related to
historical newspapers over the past decades have focused on mass
media sources and ordinary people as their primary recipients. Much
less attention has been paid to newspapers published for a more
specialized audience, e.g., those aiming at scholarly or cultural
exchange within intellectual communities much narrower in scope,
such as newspapers devoted to music criticism, arts or philosophy.
Only some few of these specialized newspapers have been digitized up
until now, but they are usually not well curated in terms of
digitization quality, data formatting, completeness, redundancy (de-
duplication), supply of metadata, and, hence, searchability. This
paper describes our approach to eliminate these drawbacks for a
major German-language newspaper resource of the Romantic Age, the
Allgemeine Musikalische Zeitung (General Music Gazette). We here
focus on a workflow that copes with a posteriori digitization
problems, inconsistent OCRing and index building for searchability.
In addition, we provide a user-friendly graphic interface to empower
content-centric access to this (and other) digital resource(s)
adopting open-source software for the purpose of Web presentation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.122}
```

}

@InProceedings{cinkova-rybicki:2020:LREC,
  author   = {Cinkova, Silvie and Rybicki, Jan},
  title   = {Stylometry in a Bilingual Setup},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {977--984},
  abstract = {The method of stylometry by most frequent words does not allow direct comparison of original texts and their translations, i.e. across languages. For instance, in a bilingual Czech-German text collection containing parallel texts (originals and translations in both directions, along with Czech and German translations from other languages), authors would not cluster across languages, since frequency word lists for any Czech texts are obviously going to be more similar to each other than to a German text, and the other way round. We have tried to come up with an interlingua that would remove the language-specific features and possibly keep the linguistically independent features of individual author signal, if they exist. We have tagged, lemmatized, and parsed each language counterpart with the corresponding language model in UDPipe, which provides a linguistic markup that is cross-lingual to a significant extent. We stripped the output of language-dependent items, but that alone did not help much. As a next step, we transformed the lemmas of both language counterparts into shared pseudolemmas based on a very crude Czech-German glossary, with a 95.6\% success. We show that, for stylometric methods based on the most frequent words, we can do without translations.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.123}
}

@InProceedings{sato-heffernan:2020:LREC,
  author   = {Sato, Yo and Heffernan, Kevin},
  title   = {Dialect Clustering with Character-Based Metrics: in Search of the Boundary of Language and Dialect},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {985--990},
  abstract = {We present in this work a universal, character-based method for representing sentences so that one can thereby calculate the distance between any two sentence pair. With a small alphabet, it can function as a proxy of phonemes, and as one of its main uses, we carry out dialect clustering: cluster a dialect/sub-language mixed corpus into sub-groups and see if they coincide with the conventional boundaries of dialects and sub-languages. By using data with multiple Japanese dialects and multiple Slavic languages, we

report how well each group clusters, in a manner to partially respond to the question of what separates languages from dialects.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.124}
}

@InProceedings{sileo-EtAl:2020:LREC,
  author   = {Sileo, Damien and Van de Cruys, Tim and Pradel, Camille and Muller, Philippe},
  title   = {DiscSense: Automated Semantic Analysis of Discourse Markers},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {991--999},
  abstract = {Using a model trained to predict discourse markers between sentence pairs, we predict plausible markers between sentence pairs with a known semantic relation (provided by existing classification datasets). These predictions allow us to study the link between discourse markers and the semantic relations annotated in classification datasets. Handcrafted mappings have been proposed between markers and discourse relations on a limited set of markers and a limited set of categories, but there exists hundreds of discourse markers expressing a wide variety of relations, and there is no consensus on the taxonomy of relations between competing discourse theories (which are largely built in a top-down fashion). By using an automatic prediction method over existing semantically annotated datasets, we provide a bottom-up characterization of discourse markers in English. The resulting dataset, named DiscSense, is publicly available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.125}
}

@InProceedings{dominguez-soler-wanner:2020:LREC,
  author   = {Dominguez, Monica and Soler, Juan and Wanner, Leo},
  title   = {ThemePro: A Toolkit for the Analysis of Thematic Progression},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1000--1007},
  abstract = {This paper introduces ThemePro, a toolkit for the automatic analysis of thematic progression. Thematic progression is relevant to natural language processing (NLP) applications dealing, among others, with discourse structure, argumentation structure, natural language generation, summarization and topic detection. A web platform demonstrates the potential of this toolkit and provides a visualization of the results including syntactic trees,

hierarchical thematicity over propositions and thematic progression over whole texts.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.126}
}

@InProceedings{jo-EtAl:2020:LREC,
  author   = {Jo, Yohan and Mayfield, Elijah and Reed, Chris and Hovy, Eduard},
  title   = {Machine-Aided Annotation for Fine-Grained Proposition Types in Argumentation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1008--1018},
  abstract = {We introduce a corpus of the 2016 U.S. presidential debates and commentary, containing 4,648 argumentative propositions annotated with fine-grained proposition types. Modern machine learning pipelines for analyzing argument have difficulty distinguishing between types of propositions based on their factuality, rhetorical positioning, and speaker commitment. Inability to properly account for these facets leaves such systems inaccurate in understanding of fine-grained proposition types. In this paper, we demonstrate an approach to annotating for four complex proposition types, namely normative claims, desires, future possibility, and reported speech. We develop a hybrid machine learning and human workflow for annotation that allows for efficient and reliable annotation of complex linguistic phenomena, and demonstrate with preliminary analysis of rhetorical strategies and structure in presidential debates. This new dataset and method can support technical researchers seeking more nuanced representations of argument, as well as argumentation theorists developing new quantitative analyses.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.127}
}

@InProceedings{chuanan-EtAl:2020:LREC,
  author   = {Chuan-An, Lin and Hung, Shyh-Shiun and Huang, Hen-Hsen and Chen, Hsin-Hsi},
  title   = {Chinese Discourse Parsing: Model and Evaluation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1019--1024},
  abstract = {Chinese discourse parsing, which aims to identify the hierarchical relationships of Chinese elementary discourse units, has not yet a consistent evaluation metric. Although Parseval is commonly used, variations of evaluation differ from three aspects: micro vs. macro F1 scores, binary vs. multiway ground truth, and

left-heavy vs. right-heavy binarization. In this paper, we first propose a neural network model that unifies a pre-trained transformer and CKY-like algorithm, and then compare it with the previous models with different evaluation scenarios. The experimental results show that our model outperforms the previous systems. We conclude that (1) the pre-trained context embedding provides effective solutions to deal with implicit semantics in Chinese texts, and (2) using multiway ground truth is helpful since different binarization approaches lead to significant differences in performance.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.128}
}

@InProceedings{long-EtAl:2020:LREC,
  author   = {Long, Wanqiu and Cai, Xinyi and Reid, James and Webber, Bonnie and Xiong, Deyi},
  title   = {Shallow Discourse Annotation for Chinese TED Talks},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1025--1032},
  abstract = {Text corpora annotated with language-related properties are an important resource for the development of Language Technology. The current work contributes a new resource for Chinese Language Technology and for Chinese-English translation, in the form of a set of TED talks (some originally given in English, some in Chinese) that have been annotated with discourse relations in the style of the Penn Discourse TreeBank, adapted to properties of Chinese text that are not present in English. The resource is currently unique in annotating discourse-level properties of planned spoken monologues rather than of written text. An inter-annotator agreement study demonstrates that the annotation scheme is able to achieve highly reliable results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.129}
}

@InProceedings{martin-poddar-upasani:2020:LREC,
  author   = {Martin, Scott and Poddar, Shivani and Upasani, Kartikeya},
  title   = {MuDoCo: Corpus for Multidomain Coreference Resolution and Referring Expression Generation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {104--111},
  abstract = {This paper proposes a new dataset, MuDoCo, composed of authored dialogs between a fictional user and a system who are given tasks to perform within six task domains. These dialogs are

given rich linguistic annotations by expert linguists for several types of reference mentions and named entity mentions, either of which can span multiple words, as well as for coreference links between mentions. The dialogs sometimes cross and blend domains, and the users exhibit complex task switching behavior such as re-initiating a previous task in the dialog by referencing the entities within it. The dataset contains a total of 8,429 dialogs with an average of 5.36 turns per dialog. We are releasing this dataset to encourage research in the field of coreference resolution, referring expression generation and identification within realistic, deep dialogs involving multiple domains. To demonstrate its utility, we also propose two baseline models for the downstream tasks: coreference resolution and referring expression generation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.13}
}

@InProceedings{olshefski-EtAl:2020:LREC,
  author   = {Olshefski, Christopher and Lugini, Luca and Singh, Ravneet and Litman, Diane and Godley, Amanda},
  title   = {The Discussion Tracker Corpus of Collaborative Argumentation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1033--1043},
  abstract = {Although NLP research on argument mining has advanced considerably in recent years, most studies draw on corpora of asynchronous and written texts, often produced by individuals. Few published corpora of synchronous, multi-party argumentation are available. The Discussion Tracker corpus, collected in high school English classes, is an annotated dataset of transcripts of spoken, multi-party argumentation. The corpus consists of 29 multi-party discussions of English literature transcribed from 985 minutes of audio. The transcripts were annotated for three dimensions of collaborative argumentation: argument moves (claims, evidence, and explanations), specificity (low, medium, high) and collaboration (e.g., extensions of and disagreements about others' ideas). In addition to providing descriptive statistics on the corpus, we provide performance benchmarks and associated code for predicting each dimension separately, illustrate the use of the multiple annotations in the corpus to improve performance via multi-task learning, and finally discuss other ways the corpus might be used to further NLP research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.130}
}

@InProceedings{sluytergthje-bourgonje-stede:2020:LREC,
  author   = {Sluyter-Gäthje, Henny and Bourgonje, Peter and Stede, Manfred},
  title   = {Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1044--1050},
  abstract  = {Shallow Discourse Parsing (SDP), the identification
of coherence relations between text spans, relies on large amounts
of training data, which so far exists only for English - any other
language is in this respect an under-resourced one. For those
languages where machine translation from English is available with
reasonable quality, MT in conjunction with annotation projection can
be an option for producing an SDP resource. In our study, we
translate the English Penn Discourse TreeBank into German and
experiment with various methods of annotation projection to arrive
at the German counterpart of the PDTB. We describe the key
characteristics of the corpus as well as some typical sources of
errors encountered during its creation. Then we evaluate the
GermanPDTB by training components for selected sub-tasks of
discourse parsing on this silver data and compare performance to the
same components when trained on the gold, original PDTB corpus.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.131}
}

@InProceedings{rasmussen-schuler:2020:LREC,
  author    = {Rasmussen, Nathan  and  Schuler, William},
  title     = {A Corpus of Encyclopedia Articles with Logical
Forms},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1051--1060},
  abstract  = {People can extract precise, complex logical meanings
from text in documents such as tax forms and game rules, but
language processing systems lack adequate training and evaluation
resources to do these kinds of tasks reliably. This paper describes
a corpus of annotated typed lambda calculus translations for
approximately 2,000 sentences in Simple English Wikipedia, which is
assumed to constitute a broad-coverage domain for precise, complex
descriptions. The corpus described in this paper contains a large
number of quantifiers and interesting scoping configurations, and is
presented specifically as a resource for quantifier scope
disambiguation systems, but also more generally as an object of
linguistic study.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.132}
}

@InProceedings{bourgonje-stede:2020:LREC,
  author    = {Bourgonje, Peter  and  Stede, Manfred},
  title     = {The Potsdam Commentary Corpus 2.2: Extending
```

Annotations for Shallow Discourse Parsing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1061--1066},
  abstract = {We present the Potsdam Commentary Corpus 2.2, a German corpus of news editorials annotated on several different levels. New in the 2.2 version of the corpus are two additional annotation layers for coherence relations following the Penn Discourse TreeBank framework. Specifically, we add relation senses to an already existing layer of discourse connectives and their arguments, and we introduce a new layer with additional coherence relation types, resulting in a German corpus that mirrors the PDTB. The aim of this is to increase usability of the corpus for the task of shallow discourse parsing. In this paper, we provide inter-annotator agreement figures for the new annotations and compare corpus statistics based on the new annotations to the equivalent statistics extracted from the PDTB.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.133}
}

@InProceedings{mohammadi-beiko-kosseim:2020:LREC,
  author   = {Mohammadi, Elham and Beiko, Timothe and Kosseim, Leila},
  title   = {On the Creation of a Corpus for Coherence Evaluation of Discursive Units},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1067--1072},
  abstract = {In this paper, we report on our experiments towards the creation of a corpus for coherence evaluation. Most corpora for textual coherence evaluation are composed of randomly shuffled sentences that focus on sentence ordering, regardless of whether the sentences were originally related by a discourse relation. To the best of our knowledge, no publicly available corpus has been designed specifically for the evaluation of coherence of known discursive units. In this paper, we focus on coherence modeling at the intra-discursive level and describe our approach to build a corpus of incoherent pairs of sentences. We experimented with a variety of corruption strategies to create synthetic incoherent pairs of discourse arguments from coherent ones. Using discourse argument pairs from the Penn Discourse Tree Bank, we generate incoherent discourse argument pairs, by swapping either their discourse connective or a discourse argument. To evaluate how incoherent the generated corpora are, we use a convolutional neural network to try to distinguish the original pairs from the corrupted ones. Results of the classifier as well as a manual inspection of

the corpora show that generating such corpora is still a challenge as the generated instances are clearly not ``incoherent enough'', indicating that more effort should be spent on developing more robust ways of generating incoherent corpora.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.134}
}

@InProceedings{desai-dakle-moldovan:2020:LREC,
  author   = {Desai, Takshak and Dakle, Parag Pravin and Moldovan, Dan},
  title   = {Joint Learning of Syntactic Features Helps Discourse Segmentation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1073--1080},
  abstract = {This paper describes an accurate framework for carrying out multi-lingual discourse segmentation with BERT (Devlin et al., 2019). The model is trained to identify segments by casting the problem as a token classification problem and jointly learning syntactic features like part-of-speech tags and dependency relations. This leads to significant improvements in performance. Experiments are performed in different languages, such as English, Dutch, German, Portuguese Brazilian and Basque to highlight the cross-lingual effectiveness of the segmenter. In particular, the model achieves a state-of-the-art F-score of 96.7 for the RST-DT corpus (Carlson et al., 2003) improving on the previous best model by 7.2\%. Additionally, a qualitative explanation is provided for how proposed changes contribute to model performance by analyzing errors made on the test data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.135}
}

@InProceedings{ruf-navarretta:2020:LREC,
  author   = {Ruf, Verena and Navarretta, Costanza},
  title   = {Creating a Corpus of Gestures and Predicting the Audience Response based on Gestures in Speeches of Donald Trump},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1081--1088},
  abstract = {Gestures are an important component of non-verbal communication. This has an increasing potential in human-computer interaction. For example, Navarretta (2017b) uses sequences of speech and pauses together with co-speech gestures produced by Barack Obama in order to predict audience response, such as applause. The aim of this study is to explore the role of speech pauses and gestures alone as predictors of audience reaction without

other types of speech information. For this work, we created a corpus of speeches held by Donald Trump before and during his time as president between 2016 and 2019. The data were transcribed with pause information and co-speech gestures were annotated as well as audience responses. Gestures and long silent pauses of the duration of at least 0.5 seconds are the input of computational models to predict audience reaction. The results of this study indicate that especially head movements and facial expressions play an important role and they confirm that gestures can to some extent be used to predict audience reaction independently of speech.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.136}
}

@InProceedings{polkov-EtAl:2020:LREC,
  author   = {Poláková, Lucie and Rysová, Kateřina and Rysová, Magdaléna and Mírovský, Jiří},
  title   = {GeCzLex: Lexicon of Czech and German Anaphoric Connectives},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1089--1096},
  abstract = {We introduce the first version of GeCzLex, an online electronic resource for translation equivalents of Czech and German discourse connectives. The lexicon is one of the outcomes of the research on anaphoricity and long-distance relations in discourse, it contains at present anaphoric connectives (ACs) for Czech and German connectives, and further their possible translations documented in bilingual parallel corpora (not necessarily anaphoric). As a basis, we use two existing monolingual lexicons of connectives: the Lexicon of Czech Discourse Connectives (CzeDLex) and the Lexicon of Discourse Markers (DiMLex) for German, interlink their relevant entries via semantic annotation of the connectives (according to the PDTB 3 sense taxonomy) and statistical information of translation possibilities from the Czech and German parallel data of the InterCorp project. The lexicon is, as far as we know, the first bilingual inventory of connectives with linkage on the level of individual entries, and a first attempt to systematically describe devices engaged in long-distance, non-local discourse coherence. The lexicon is freely available under the Creative Commons License.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.137}
}

@InProceedings{das-EtAl:2020:LREC,
  author   = {Das, Debopam and Stede, Manfred and Ghosh, Soumya Sankar and Chatterjee, Lahari},
  title   = {DiMLex-Bangla: A Lexicon of Bangla Discourse Connectives},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1097--1102},
  abstract  = {We present DiMLex-Bangla, a newly developed lexicon
of discourse connectives in Bangla. The lexicon, upon completion of
its first version, contains 123 Bangla connective entries, which are
primarily compiled from the linguistic literature and translation of
English discourse connectives. The lexicon compilation is later
augmented by adding more connectives from a currently developed
corpus, called the Bangla RST Discourse Treebank (Das and Stede,
2018). DiMLex-Bangla provides information on syntactic categories of
Bangla connectives, their discourse semantics and non-connective
uses (if any). It uses the format of the German connective lexicon
DiMLex (Stede and Umbach, 1998), which provides a cross-
linguistically applicable XML schema. The resource is the first of
its kind in Bangla, and is freely available for use in studies on
discourse structure and computational applications.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.138}
}

@InProceedings{knaebel-stede:2020:LREC,
  author    = {Knaebel, Rene  and  Stede, Manfred},
  title     = {Semi-Supervised Tri-Training for Explicit Discourse
Argument Expansion},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1103--1109},
  abstract  = {This paper describes a novel application of semi-
supervision for shallow discourse parsing. We use a neural approach
for sequence tagging and focus on the extraction of explicit
discourse arguments. First, additional unlabeled data is prepared
for semi-supervised learning. From this data, weak annotations are
generated in a first setting and later used in another setting to
study performance differences. In our studies, we show an increase
in the performance of our models that ranges between 2-10\% F1
score. Further, we give some insights to the generated discourse
annotations and compare the developed additional relations with the
training relations. We release this new dataset of explicit
discourse arguments to enable the training of large statistical
models.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.139}
}

@InProceedings{xiang-EtAl:2020:LREC1,
  author    = {Xiang, Rong  and  Long, Yunfei  and  Wan, Mingyu  and
Gu, Jinghang  and  Lu, Qin  and  Huang, Chu-Ren},
  title     = {Affection Driven Neural Networks for Sentiment
Analysis},
```

  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {112--119},
  abstract = {Deep neural network models have played a critical role in sentiment analysis with promising results in the recent decade. One of the essential challenges, however, is how external sentiment knowledge can be effectively utilized. In this work, we propose a novel affection-driven approach to incorporating affective knowledge into neural network models. The affective knowledge is obtained in the form of a lexicon under the Affect Control Theory (ACT), which is represented by vectors of three-dimensional attributes in Evaluation, Potency, and Activity (EPA). The EPA vectors are mapped to an affective influence value and then integrated into Long Short-term Memory (LSTM) models to highlight affective terms. Experimental results show a consistent improvement of our approach over conventional LSTM models by 1.0\% to 1.5\% in accuracy on three large benchmark datasets. Evaluations across a variety of algorithms have also proven the effectiveness of leveraging affective terms for deep model enhancement.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.14}
}

@InProceedings{chinnappa-palmer-blanco:2020:LREC,
  author   = {Chinnappa, Dhivya and Palmer, Alexis and Blanco, Eduardo},
  title   = {WikiPossessions: Possession Timeline Generation as an Evaluation Benchmark for Machine Reading Comprehension of Long Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1110--1117},
  abstract = {This paper presents WikiPossessions, a new benchmark corpus for the task of temporally-oriented possession (TOP), or tracking objects as they change hands over time. We annotate Wikipedia articles for 90 different well-known artifacts paintings, diamonds, and archaeological artifacts), producing 799 artifact-possessor relations with associated attributes. For each article, we also produce a full possession timeline. The full version of the task combines straightforward entity-relation extraction with complex temporal reasoning, as well as verification of textual support for the relevant types of knowledge. Specifically, to complete the full TOP task for a given article, a system must do the following: a) identify possessors; b) anchor possessors to times/ events; c) identify temporal relations between each temporal anchor and the possession relation it corresponds to; d) assign certainty scores to each possessor and each temporal relation; and e) assemble

```
individual possession events into a global possession timeline. In
addition to the corpus, we release evaluation scripts and a baseline
model for the task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.140}
}

@InProceedings{westera-mayol-rohde:2020:LREC,
  author    = {Westera, Matthijs  and  Mayol, Laia  and  Rohde,
Hannah},
  title     = {TED-Q: TED Talks and the Questions they Evoke},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1118--1127},
  abstract  = {We present a new dataset of TED-talks annotated with
the questions they evoke and, where available, the answers to these
questions. Evoked questions represent a hitherto mostly unexplored
type of linguistic data, which promises to open up important new
lines of research, especially related to the Question Under
Discussion (QUD)-based approach to discourse structure. In this
paper we introduce the method and open the first installment of our
data to the public. We summarize and explore the current dataset,
illustrate its potential by providing new evidence for the relation
between predictability and implicitness -- capitalizing on the
already existing PDTB-style annotations for the texts we use -- and
outline its potential for future research. The dataset should be of
interest, at its current scale, to researchers on formal and
experimental pragmatics, discourse coherence, information structure,
discourse expectations and processing. Our data-gathering procedure
is designed to scale up, relying on crowdsourcing by non-expert
annotators, with its utility for Natural Language Processing in mind
(e.g., dialogue systems, conversational question answering).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.141}
}

@InProceedings{mrovsk-polkov-synkov:2020:LREC,
  author    = {Mírovský, Jiří  and  Poláková, Lucie  and  Synková,
Pavlína},
  title     = {CzeDLex 0.6 and its Representation in the PML-TQ},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1128--1134},
  abstract  = {CzeDLex is an electronic lexicon of Czech discourse
connectives with its data coming from a large treebank annotated
with discourse relations. Its new version CzeDLex 0.6 (as compared
with the previous version 0.5, which was published in 2017) is
significantly larger with respect to manually processed entries.
```

```
Also, its structure has been modified to allow for primary
connectives to appear with multiple entries for a single discourse
sense. The lexicon comes in several formats, being both human and
machine readable, and is available for searching in PML Tree Query,
a user-friendly and powerful search tool for all kinds of
linguistically annotated treebanks. The main purpose of this paper/
demo is to present the new version of the lexicon and to demonstrate
possibilities of mining various types of information from the
lexicon using PML Tree Query; we present several examples of search
queries over the lexicon data along with their results. The new
version of the lexicon, CzeDLex~0.6, is available on-line and was
officially released in December 2019 under the Creative Commons
License.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.142}
}

@InProceedings{egawa-morio-fujita:2020:LREC,
  author    = {Egawa, Ryo  and  Morio, Gaku  and  Fujita,
Katsuhide},
  title     = {Corpus for Modeling User Interactions in Online
Persuasive Discussions},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1135--1141},
  abstract  = {Persuasions are common in online arguments such as
discussion forums. To analyze persuasive strategies, it is important
to understand how individuals construct posts and comments based on
the semantics of the argumentative components. In addition to
understanding how we construct arguments, understanding how a user
post interacts with other posts (i.e., argumentative inter-post
relation) still remains a challenge. Therefore, in this study, we
developed a novel annotation scheme and corpus that capture both
user-generated inner-post arguments and inter-post relations between
users in ChangeMyView, a persuasive forum. Our corpus consists of
arguments with 4612 elementary units (EUs) (i.e., propositions),
2713 EU-to-EU argumentative relations, and 605 inter-post
argumentative relations in 115 threads. We analyzed the annotated
corpus to identify the characteristics of online persuasive
arguments, and the results revealed persuasive documents have more
claims than non-persuasive ones and different interaction patterns
among persuasive and non-persuasive documents. Our corpus can be
used as a resource for analyzing persuasiveness and training an
argument mining system to identify and extract argument structures.
The annotated corpus and annotation guidelines have been made
publicly available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.143}
}

@InProceedings{wilkens-todirascu:2020:LREC,
  author    = {Wilkens, Rodrigo  and  Todirascu, Amalia},
```

```
  title     = {Simplifying Coreference Chains for Dyslexic
Children},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1142--1151},
  abstract  = {We present a work aiming to generate adapted content
for dyslexic children for French, in the context of the ALECTOR
project. Thus, we developed a system to transform the texts at the
discourse level. This system modifies the coreference chains, which
are markers of text cohesion, by using rules. These rules were
designed following a careful study of coreference chains in both
original texts and its simplified versions. Moreover, in order to
define reliable transformation rules, we analysed several
coreference properties as well as the concurrent simplification
operations in the aligned texts. This information is coded together
with a coreference resolution system and a text rewritten tool in
the proposed system, which comprise a coreference module specialised
in written text and seven text transformation operations. The
evaluation of the system firstly focused on check the simplification
by manual validation of three judges. These errors were grouped into
five classes that combined can explain 93\% of the errors. The
second evaluation step consisted of measuring the simplification
perception by 23 judges, which allow us to measure the
simplification impact of the proposed rules.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.144}
}

@InProceedings{kishimoto-murawaki-kurohashi:2020:LREC,
  author    = {Kishimoto, Yudai  and  Murawaki, Yugo  and
Kurohashi, Sadao},
  title     = {Adapting BERT to Implicit Discourse Relation
Classification with a Focus on Discourse Connectives},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1152--1158},
  abstract  = {BERT, a neural network-based language model pre-
trained on large corpora, is a breakthrough in natural language
processing, significantly outperforming previous state-of-the-art
models in numerous tasks. However, there have been few reports on
its application to implicit discourse relation classification, and
it is not clear how BERT is best adapted to the task. In this paper,
we test three methods of adaptation. (1) We perform additional pre-
training on text tailored to discourse classification. (2) In
expectation of knowledge transfer from explicit discourse relations
to implicit discourse relations, we add a task named explicit
connective prediction at the additional pre-training step. (3) To
```

```
exploit implicit connectives given by treebank annotators, we add a
task named implicit connective prediction at the fine-tuning step.
We demonstrate that these three techniques can be combined
straightforwardly in a single training pipeline. Through
comprehensive experiments, we found that the first and second
techniques provide additional gain while the last one did not.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.145}
}

@InProceedings{barbedette-eshkoltaravella:2020:LREC,
  author    = {Barbedette, Angèle  and  Eshkol-Taravella, Iris},
  title     = {What Speakers really Mean when they Ask Questions:
Classification of Intentions with a Supervised Approach},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1159--1166},
  abstract  = {This paper focuses on the automatic detection of
hidden intentions of speakers in questions asked during meals. Our
corpus is composed of a set of transcripts of spontaneous oral
conversations from ESLO's corpora. We suggest a typology of these
intentions based on our research work and the exploration and
annotation of the corpus, in which we define two "explicit"
categories (request for agreement and request for information) and
three "implicit" categories (opinion, will and doubt). We implement
a supervised automatic classification model based on annotated data
and selected linguistic features and we evaluate its results and
performances. We finally try to interpret these results by looking
more deeply and specifically into the predictions of the algorithm
and the features it used. There are many motivations for this work
which are part of ongoing challenges such as opinion analysis, irony
detection or the development of conversational agents.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.146}
}

@InProceedings{farzana-valizadeh-parde:2020:LREC,
  author    = {Farzana, Shahla  and  Valizadeh, Mina  and  Parde,
Natalie},
  title     = {Modeling Dialogue in Conversational Cognitive Health
Screening Interviews},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1167--1177},
  abstract  = {Automating straightforward clinical tasks can reduce
workload for healthcare professionals, increase accessibility for
geographically-isolated patients, and alleviate some of the economic
burdens associated with healthcare. A variety of preliminary
```

screening procedures are potentially suitable for automation, and one such domain that has remained underexplored to date is that of structured clinical interviews. A task-specific dialogue agent is needed to automate the collection of conversational speech for further (either manual or automated) analysis, and to build such an agent, a dialogue manager must be trained to respond to patient utterances in a manner similar to a human interviewer. To facilitate the development of such an agent, we propose an annotation schema for assigning dialogue act labels to utterances in patient-interviewer conversations collected as part of a clinically-validated cognitive health screening task. We build a labeled corpus using the schema, and show that it is characterized by high inter-annotator agreement. We establish a benchmark dialogue act classification model for the corpus, thereby providing a proof of concept for the proposed annotation schema. The resulting dialogue act corpus is the first such corpus specifically designed to facilitate automated cognitive health screening, and lays the groundwork for future exploration in this area.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.147}
}

@InProceedings{straton-jang-ng:2020:LREC,
  author   = {Straton, Nadiya and Jang, Hyeju and Ng, Raymond},
  title   = {Stigma Annotation Scheme and Stigmatized Language Detection in Health-Care Discussions on Social Media},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1178--1190},
  abstract = {Much research has been done within the social sciences on the interpretation and influence of stigma on human behaviour and health, which result in out-of-group exclusion, distancing, cognitive separation, status loss, discrimination, in-group pressure, and often lead to disengagement, non-adherence to treatment plan, and prescriptions by the doctor. However, little work has been conducted on computational identification of stigma in general and in social media discourse in particular. In this paper, we develop the annotation scheme and improve the annotation process for stigma identification, which can be applied to other health-care domains. The data from pro-vaccination and anti-vaccination discussion groups are annotated by trained annotators who have professional background in social science and health-care studies, therefore the group can be considered experts on the subject in comparison to non-expert crowd. Amazon MTurk annotators is another group of annotator with no knowledge on their education background, they are initially treated as non-expert crowd on the subject matter of stigma. We analyze the annotations with visualisation techniques, features from LIWC (Linguistic Inquiry and Word Count) list and make prediction based on bi-grams with traditional and deep learning models. Data augmentation method and application of CNN show high performance accuracy in comparison to other models. Success of the

rigorous annotation process on identifying stigma is reconfirmed by achieving high prediction rate with CNN.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.148}
}

@InProceedings{dhanwal-EtAl:2020:LREC,
  author   = {Dhanwal, Swapnil and Dutta, Hritwik and Nankani, Hitesh and Shrivastava, Nilay and Kumar, Yaman and Li, Junyi Jessy and Mahata, Debanjan and Gosangi, Rakesh and Zhang, Haimin and Shah, Rajiv Ratn and Stent, Amanda},
  title   = {An Annotated Dataset of Discourse Modes in Hindi Stories},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1191--1196},
  abstract = {In this paper, we present a new corpus consisting of sentences from Hindi short stories annotated for five different discourse modes argumentative, narrative, descriptive, dialogic and informative. We present a detailed account of the entire data collection and annotation processes. The annotations have a very high inter-annotator agreement (0.87 k-alpha). We analyze the data in terms of label distributions, part of speech tags, and sentence lengths. We characterize the performance of various classification algorithms on this dataset and perform ablation studies to understand the nature of the linguistic models suitable for capturing the nuances of the embedded discourse structures in the presented corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.149}
}

@InProceedings{bhattasali-EtAl:2020:LREC,
  author   = {Bhattasali, Shohini and Brennan, Jonathan and Luh, Wen-Ming and Franzluebbers, Berta and Hale, John},
  title   = {The Alice Datasets: fMRI \& EEG Observations of Natural Language Comprehension},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {120--125},
  abstract = {The Alice Datasets are a set of datasets based on magnetic resonance data and electrophysiological data, collected while participants heard a story in English. Along with the datasets and the text of the story, we provide a variety of different linguistic and computational measures ranging from prosodic predictors to predictors capturing hierarchical syntactic information. These ecologically valid datasets can be easily reused to replicate prior work and to test new hypotheses about natural

```
language comprehension in the brain.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.15}
}

@InProceedings{shavarani-sekine:2020:LREC,
  author    = {Shavarani, Hassan S.  and  Sekine, Satoshi},
  title     = {Multi-class Multilingual Classification of Wikipedia
Articles Using Extended Named Entity Tag Set},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1197--1201},
  abstract  = {Wikipedia is a great source of general world
knowledge which can guide NLP models better understand their
motivation to make predictions. Structuring Wikipedia is the initial
step towards this goal which can facilitate fine-grain
classification of articles. In this work, we introduce the Shinra 5-
Language Categorization Dataset (SHINRA-5LDS), a large multi-lingual
and multi-labeled set of annotated Wikipedia articles in Japanese,
English, French, German, and Farsi using Extended Named Entity (ENE)
tag set. We evaluate the dataset using the best models provided for
ENE label set classification and show that the currently available
classification models struggle with large datasets using fine-
grained tag sets.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.150}
}

@InProceedings{moudjari-akliastouati-benamara:2020:LREC,
  author    = {Moudjari, Leila  and  Akli-Astouati, Karima  and
Benamara, Farah},
  title     = {An Algerian Corpus and an Annotation Platform for
Opinion and Emotion Analysis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1202--1210},
  abstract  = {In this paper, we address the lack of resources for
opinion and emotion analysis related to North African dialects,
targeting Algerian dialect. We present TWIFIL (TWItter proFILing) a
collaborative annotation platform for crowdsourcing annotation of
tweets at different levels of granularity. The plateform allowed the
creation of the largest Algerian dialect dataset annotated for both
sentiment (9,000 tweets), emotion (about 5,000 tweets) and extra-
linguistic information including author profiling (age and gender).
The annotation resulted also in the creation of the largest Algerien
dialect subjectivity lexicon of about 9,000 entries which can
constitute a valuable resources for the development of future NLP
applications for Algerian dialect. To test the validity of the
```

dataset, a set of deep learning experiments were conducted to classify a given tweet as positive, negative or neutral. We discuss our results and provide an error analysis to better identify classification errors.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.151}
}

@InProceedings{slovikovskaya-attardi:2020:LREC,
  author   = {Slovikovskaya, Valeriya and Attardi, Giuseppe},
  title   = {Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1211--1218},
  abstract = {Transformer models, trained and publicly released over the last couple of years, have proved effective in many NLP tasks. We wished to test their usefulness in particular on the stance detection task. We performed experiments on the data from the Fake News Challenge Stage 1 (FNC-1). We were indeed able to improve the reported SotA on the challenge, by exploiting the generalization power of large language models based on Transformer architecture. Specifically (1) we improved the FNC-1 best performing model adding BERT sentence embedding of input sequences as a model feature, (2) we fine-tuned BERT, XLNet, and RoBERTa transformers on FNC-1 extended dataset and obtained state-of-the-art results on FNC-1 task.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.152}
}

@InProceedings{ginev-miller:2020:LREC,
  author   = {Ginev, Deyan and Miller, Bruce R},
  title   = {Scientific Statement Classification over arXiv.org},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1219--1226},
  abstract = {We introduce a new classification task for scientific statements and release a large-scale dataset for supervised learning. Our resource is derived from a machine-readable representation of the arXiv.org collection of preprint articles. We explore fifty author-annotated categories and empirically motivate a task design of grouping 10.5 million annotated paragraphs into thirteen classes. We demonstrate that the task setup aligns with known success rates from the state of the art, peaking at a 0.91 F1-score via a BiLSTM encoder-decoder model. Additionally, we introduce a lexeme serialization for mathematical formulas, and observe that context-aware models could improve when also trained on the symbolic

```
modality. Finally, we discuss the limitations of both data and task
design, and outline potential directions towards increasingly
complex models of scientific discourse, beyond isolated
statements.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.153}
}

@InProceedings{dias-paraboni:2020:LREC,
  author    = {Dias, Rafael  and  Paraboni, Ivandré},
  title     = {Cross-domain Author Gender Classification in
Brazilian Portuguese},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1227--1234},
  abstract  = {Author profiling models predict demographic
characteristics of a target author based on the text that they have
written. Systems of this kind will often follow a single-domain
approach, in which the model is trained from a corpus of labelled
texts in a given domain, and it is subsequently validated against a
test corpus built from precisely the same domain. Although single-
domain settings are arguably ideal, this strategy gives rise to the
question of how to proceed when no suitable training corpus (i.e., a
corpus that matches the test domain) is available. To shed light on
this issue, this paper discusses a cross-domain gender
classification task based on four domains (Facebook, crowd sourced
opinions, Blogs and \mbox{E-gov} requests) in the Brazilian
Portuguese language. A number of simple gender classification models
using word- and psycholinguistics-based features alike are
introduced, and their results are compared in two kinds of cross-
domain setting: first, by making use of a single text source as
training data for each task, and subsequently by combining multiple
sources. Results confirm previous findings related to the effects of
corpus size and domain similarity in English, and pave the way for
further studies in the field.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.154}
}

@InProceedings{tuggener-EtAl:2020:LREC,
  author    = {Tuggener, Don  and  von Däniken, Pius  and  Peetz,
Thomas  and  Cieliebak, Mark},
  title     = {LEDGAR: A Large-Scale Multi-label Corpus for Text
Classification of Legal Provisions in Contracts},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1235--1241},
  abstract  = {We present LEDGAR, a multilabel corpus of legal
```

provisions in contracts. The corpus was crawled and scraped from the public domain (SEC filings) and is, to the best of our knowledge, the first freely available corpus of its kind. Since the corpus was constructed semi-automatically, we apply and discuss various approaches to noise removal. Due to the rather large labelset of over 12'000 labels annotated in almost 100'000 provisions in over 60'000 contracts, we believe the corpus to be of interest for research in the field of Legal NLP, (large-scale or extreme) text classification, as well as for legal studies. We discuss several methods to sample subcopora from the corpus and implement and evaluate different automatic classification approaches. Finally, we perform transfer experiments to evaluate how well the classifiers perform on contracts stemming from outside the corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.155}
}

@InProceedings{rodier-carter:2020:LREC,
  author   = {Rodier, Simon and Carter, Dave},
  title   = {Online Near-Duplicate Detection of News Articles},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1242--1249},
  abstract = {Near-duplicate documents are particularly common in news media corpora. Editors often update wirefeed articles to address space constraints in print editions or to add local context; journalists often lightly modify previous articles with new information or minor corrections. Near-duplicate documents have potentially significant costs, including bloating corpora with redundant information (biasing techniques built upon such corpora) and requiring additional human and computational analytic resources for marginal benefit. Filtering near-duplicates out of a collection is thus important, and is particularly challenging in applications that require them to be filtered out in real-time with high precision. Previous near-duplicate detection methods typically work offline to identify all near-duplicate pairs in a set of documents. We propose an online system which flags a near-duplicate document by finding its most likely original. This system adapts the shingling algorithm proposed by Broder (1997), and we test it on a challenging dataset of web-based news articles. Our online system presents state-of-the-art F1-scores, and can be tuned to trade precision for recall and vice-versa. Given its performance and online nature, our method can be used in many real-world applications. We present one such application, filtering near-duplicates to improve productivity of human analysts in a situational awareness tool.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.156}
}

@InProceedings{hirao-EtAl:2020:LREC,
  author   = {Hirao, Reo and Arai, Mio and Shimanaka, Hiroki and Katsumata, Satoru and Komachi, Mamoru},

```
  title     = {Automated Essay Scoring System for Nonnative Japanese
Learners},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1250--1257},
  abstract  = {In this study, we created an automated essay scoring
(AES) system for nonnative Japanese learners using an essay dataset
with annotations for a holistic score and multiple trait scores,
including content, organization, and language scores. In particular,
we developed AES systems using two different approaches: a feature-
based approach and a neural-network-based approach. In the former
approach, we used Japanese-specific linguistic features, including
character-type features such as “kanji” and “hiragana.” In the
latter approach, we used two models: a long short-term memory (LSTM)
model (Hochreiter and Schmidhuber, 1997) and a bidirectional encoder
representations from transformers (BERT) model (Devlin et al.,
2019), which achieved the highest accuracy in various natural
language processing tasks in 2018. Overall, the BERT model achieved
the best root mean squared error and quadratic weighted kappa
scores. In addition, we analyzed the robustness of the outputs of
the BERT model. We have released and shared this system to
facilitate further research on AES for Japanese as a second language
learners.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.157}
}

@InProceedings{neerbek-EtAl:2020:LREC,
  author    = {Neerbek, Jan  and  Eskildsen, Morten  and  Dolog,
Peter  and  Assent, Ira},
  title     = {A Real-World Data Resource of Complex Sensitive
Sentences Based on Documents from the Monsanto Trial},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1258--1267},
  abstract  = {In this work we present a corpus for the evaluation
of sensitive information detection approaches that addresses the
need for real world sensitive information for empirical studies. Our
sentence corpus contains different notions of complex sensitive
information that correspond to different aspects of concern in a
current trial of the Monsanto company. This paper describes the
annotations process, where we both employ human annotators and
furthermore create automatically inferred labels regarding
technical, legal and informal communication within and with
employees of Monsanto, drawing on a classification of documents by
lawyers involved in the Monsanto court case. We release corpus of
high quality sentences and parse trees with these two types of
```

labels on sentence level. We characterize the sensitive information via several representative sensitive information detection models, in particular both keyword-based (n-gram) approaches and recent deep learning models, namely, recurrent neural networks (LSTM) and recursive neural networks (RecNN). Data and code are made publicly available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.158}
}

@InProceedings{lazaridou-EtAl:2020:LREC,
  author   = {Lazaridou, Konstantina and Löser, Alexander and Mestre, Maria and Naumann, Felix},
  title   = {Discovering Biased News Articles Leveraging Multiple Human Annotations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1268--1277},
  abstract = {Unbiased and fair reporting is an integral part of ethical journalism. Yet, political propaganda and one-sided views can be found in the news and can cause distrust in media. Both accidental and deliberate political bias affect the readers and shape their views. We contribute to a trustworthy media ecosystem by automatically identifying politically biased news articles. We introduce novel corpora annotated by two communities, i.e., domain experts and crowd workers, and we also consider automatic article labels inferred by the newspapers' ideologies. Our goal is to compare domain experts to crowd workers and also to prove that media bias can be detected automatically. We classify news articles with a neural network and we also improve our performance in a self-supervised manner.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.159}
}

@InProceedings{mikhalkova-EtAl:2020:LREC,
  author   = {Mikhalkova, Elena and Protasov, Timofei and Sokolova, Polina and Bashmakova, Anastasiia and Drozdova, Anastasiia},
  title   = {Modelling Narrative Elements in a Short Story: A Study on Annotation Schemes and Guidelines},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {126--132},
  abstract = {Text-processing algorithms that annotate main components of a story-line are presently in great need of corpora and well-agreed annotation schemes. The Text World Theory of cognitive linguistics offers a model that generalizes a narrative

```
structure in the form of world building elements (characters, time
and space) as well as text worlds themselves and switches between
them. We have conducted a survey on how text worlds and their
elements are annotated in different projects and proposed our own
annotation scheme and instructions. We tested them, first, on the
science fiction story ``We Can Remember It for You Wholesale'' by
Philip K. Dick. Then we corrected the guidelines and added computer
annotation of verb forms with the purpose to get a higher raters'
agreement and tested them again on the short story ``The Gift of the
Magi'' by O. Henry. As a result, the agreement among the three
raters has risen. With due revision and tests, our annotation scheme
and guidelines can be used for annotating narratives in corpora of
literary texts, criminal evidence, teaching materials, quests,
etc.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.16}
}

@InProceedings{gonalooliveira-clemncio-alves:2020:LREC,
  author    = {Gonçalo Oliveira, Hugo  and  Clemêncio, André  and
Alves, Ana},
  title     = {Corpora and Baselines for Humour Recognition in
Portuguese},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1278--1285},
  abstract  = {Having in mind the lack of work on the automatic
recognition of verbal humour in Portuguese, a topic connected with
fluency in a natural language, we describe the creation of three
corpora, covering two styles of humour and four sources of non-
humorous text, that may be used for related studies. We then report
on some experiments where the created corpora were used for training
and testing computational models that exploit content and linguistic
features for humour recognition. The obtained results helped us
taking some conclusions about this challenge and may be seen as
baselines for those willing to tackle it in the future, using the
same corpora.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.160}
}

@InProceedings{vandermeulen-reijnierse:2020:LREC,
  author    = {van der Meulen, Marten  and  Reijnierse, W. Gudrun},
  title     = {FactCorp: A Corpus of Dutch Fact-checks and its
Multiple Usages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1286--1292},
```

```
  abstract  = {Fact-checking information before publication has long
been a core task for journalists, but recent times have seen the
emergence of dedicated news items specifically aimed at fact-
checking after publication. This relatively new form of fact-
checking receives a fair amount of attention from academics, with
current research focusing mostly on journalists' motivations for
publishing post-hoc fact-checks, the effects of fact-checking on the
perceived accuracy of false claims, and the creation of
computational tools for automatic fact-checking. In this paper, we
propose to study fact-checks from a corpus linguistic perspective.
This will enable us to gain insight in the scope and contents of
fact-checks, to investigate what fact-checks can teach us about the
way in which science appears (incorrectly) in the news, and to see
how fact-checks behave in the science communication landscape. We
report on the creation of FactCorp, a 1,16 million-word corpus
containing 1,974 fact-checks from three major Dutch newspapers. We
also present results of several exploratory analyses, including a
rhetorical moves analysis, a qualitative content elements analysis,
and keyword analyses. Through these analyses, we aim to demonstrate
the wealth of possible applications that FactCorp allows, thereby
stressing the importance of creating such resources.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.161}
}

@InProceedings{ortmann-dipper:2020:LREC,
  author    = {Ortmann, Katrin  and  Dipper, Stefanie},
  title     = {Automatic Orality Identification in Historical
Texts},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1293--1302},
  abstract  = {Independently of the medial representation (written/
spoken), language can exhibit characteristics of conceptual orality
or literacy, which mainly manifest themselves on the lexical or
syntactic level. In this paper we aim at automatically identifying
conceptually-oral historical texts, with the ultimate goal of
gaining knowledge about spoken data of historical time stages. We
apply a set of general linguistic features that have been proven to
be effective for the classification of modern language data to
historical German texts from various registers. Many of the features
turn out to be equally useful in determining the conceptuality of
historical data as they are for modern data, especially the
frequency of different types of pronouns and the ratio of verbs to
nouns. Other features like sentence length, particles or
interjections point to peculiarities of the historical data and
reveal problems with the adoption of a feature set that was
developed on modern language data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.162}
}
```

```
@InProceedings{song-EtAl:2020:LREC1,
  author    = {Song, Xingyi  and  Downs, Johnny  and  Velupillai, Sumithra  and  Holden, Rachel  and  Kikoler, Maxim  and  Bontcheva, Kalina  and  Dutta, Rina  and  Roberts, Angus},
  title     = {Using Deep Neural Networks with Intra- and Inter-Sentence Context to Classify Suicidal Behaviour},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1303--1310},
  abstract  = {Identifying statements related to suicidal behaviour in psychiatric electronic health records (EHRs) is an important step when modeling that behaviour, and when assessing suicide risk. We apply a deep neural network based classification model with a lightweight context encoder, to classify sentence level suicidal behaviour in EHRs. We show that incorporating information from sentences to left and right of the target sentence significantly improves classification accuracy. Our approach achieved the best performance when classifying suicidal behaviour in Autism Spectrum Disorder patient records. The results could have implications for suicidality research and clinical surveillance.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.163}
}

@InProceedings{mohamed-ha:2020:LREC,
  author    = {Mohamed, Emad  and  Ha, Le An},
  title     = {A First Dataset for Film Age Appropriateness Investigation},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1311--1317},
  abstract  = {Film age appropriateness classification is an important problem with a significant societal impact that has so far been out of the interest of Natural Language Processing and Machine Learning researchers. To this end, we have collected a corpus of 17000 films along with their age ratings. We use the textual contents in an experiment to predict the correct age classification for the United States (G, PG, PG-13, R and NC-17) and the United Kingdom (U, PG, 12A, 15, 18 and R18). Our experiments indicate that gradient boosting machines beat FastText and various Deep Learning architectures. We reach an overall accuracy of 79.3\% for the US ratings compared to a projected super human accuracy of 84\%. For the UK ratings, we reach an overall accuracy of 65.3\% (UK) compared to a projected super human accuracy of 80.0\%.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.164}
}
```

```
@InProceedings{elhaj:2020:LREC,
  author    = {El-Haj, Mahmoud},
  title     = {Habibi - a multi Dialect multi National Arabic Song
Lyrics Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1318--1326},
  abstract  = {This paper introduces Habibi the first Arabic Song
Lyrics corpus. The corpus comprises more than 30,000 Arabic song
lyrics in 6 Arabic dialects for singers from 18 different Arabic
countries. The lyrics are segmented into more than 500,000 sentences
(song verses) with more than 3.5 million words. I provide the corpus
in both comma separated value (csv) and annotated plain text (txt)
file formats. In addition, I converted the csv version into
JavaScript Object Notation (json) and eXtensible Markup Language
(xml) file formats. To experiment with the corpus I run extensive
binary and multi-class experiments for dialect and country-of-origin
identification. The identification tasks include the use of several
classical machine learning and deep learning models utilising
different word embeddings. For the binary dialect identification
task the best performing classifier achieved a testing accuracy of
93\%. This was achieved using a word-based Convolutional Neural
Network (CNN) utilising a Continuous Bag of Words (CBOW) word
embeddings model. The results overall show all classical and deep
learning models to outperform our baseline, which demonstrates the
suitability of the corpus for both dialect and country-of-origin
identification tasks. I am making the corpus and the trained CBOW
word embeddings freely available for research purposes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.165}
}

@InProceedings{shafaei-EtAl:2020:LREC,
  author    = {Shafaei, Mahsa  and  Safi Samghabadi, Niloofar  and
Kar, Sudipta  and  Solorio, Thamar},
  title     = {Age Suitability Rating: Predicting the MPAA Rating
Based on Movie Dialogues},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1327--1335},
  abstract  = {Movies help us learn and inspire societal change. But
they can also contain objectionable content that negatively affects
viewers' behaviour, especially children. In this paper, our goal is
to predict the suitability of movie content for children and young
adults based on scripts. The criterion that we use to measure
suitability is the MPAA rating that is specifically designed for
this purpose. We create a corpus for movie MPAA ratings and propose
```

```
an RNN based architecture with attention that jointly models the
genre and the emotions in the script to predict the MPAA rating. We
achieve 81\% weighted F1-score for the classification model that
outperforms the traditional machine learning method by 7\%.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.166}
}

@InProceedings{alkhereyf-rambow:2020:LREC,
  author    = {Alkhereyf, Sakhar  and  Rambow, Owen},
  title     = {Email Classification Incorporating Social Networks
and Thread Structure},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1336--1345},
  abstract  = {Existing methods for different document
classification tasks in the context of social networks typically
only capture the semantics of texts, while ignoring the users who
exchange the text and the network they form. However, some work has
shown that incorporating the social network information in addition
to information from language is effective for various NLP
applications including sentiment analysis, inferring user
attributes, and predicting inter-personal relations. In this paper,
we present an empirical study of email classification into
``Business'' and ``Personal'' categories. We represent the email
communication using various graph structures. As features, we use
both the textual information from the email content and social
network information from the communication graphs. We also model the
thread structure for emails. We focus on detecting personal emails,
and we evaluate our methods on two corpora, only one of which we
train on. The experimental results reveal that incorporating social
network information improves over the performance of an approach
based on textual information only. The results also show that
considering the thread structure of emails improves the performance
further. Furthermore, our approach improves over a state-of-the-art
baseline which uses node embeddings based on both lexical and social
network information.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.167}
}

@InProceedings{tseng-EtAl:2020:LREC,
  author    = {Tseng, Yuen-Hsien  and  Wu, Wun-Syuan  and  Chang,
Chia-Yueh  and  Chen, Hsueh-Chih  and  Hsu, Wei-Lun},
  title     = {Development and Validation of a Corpus for Machine
Humor Comprehension},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {1346--1352},
  abstract  = {This work developed a Chinese humor corpus containing
3,365 jokes collected from over 40 sources. Each joke was labeled
with five levels of funniness, eight skill sets of humor, and six
dimensions of intent by only one annotator. To validate the manual
labels, we trained SVM (Support Vector Machine) and BERT
(Bidirectional Encoder Representations from Transformers) with half
of the corpus (labeled by one annotator) to predict the skill and
intent labels of the other half (labeled by the other annotator).
Based on two assumptions that a valid manually labeled corpus should
follow, our results showed the validity for the skill and intent
labels. As to the funniness label, the validation results showed
that the correlation between the corpus label and user feedback
rating is marginal, which implies that the funniness level is a
harder annotation problem to be solved. The contribution of this
work is two folds: 1) a Chinese humor corpus is developed with
labels of humor skills, intents, and funniness, which allows
machines to learn more intricate humor framing, effect, and amusing
level to predict and respond in proper context (https://github.com/
SamTseng/Chinese\_Humor\_MultiLabeled). 2) An approach to verify
whether a minimum human labeled corpus is valid or not, which
facilitates the validation of low-resource corpora.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.168}
}

@InProceedings{gala-EtAl:2020:LREC,
  author    = {Gala, Núria  and  Tack, Anaïs  and  Javourey-Drevet,
Ludivine  and  François, Thomas  and  Ziegler, Johannes C.},
  title     = {Alector: A Parallel Corpus of Simplified French Texts
with Alignments of Misreadings by Poor and Dyslexic Readers},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1353--1361},
  abstract  = {In this paper, we present a new parallel corpus
addressed to researchers, teachers, and speech therapists interested
in text simplification as a means of alleviating difficulties in
children learning to read. The corpus is composed of excerpts drawn
from 79 authentic literary (tales, stories) and scientific
(documentary) texts commonly used in French schools for children
aged between 7 to 9 years old. The excerpts were manually simplified
at the lexical, morpho-syntactic, and discourse levels in order to
propose a parallel corpus for reading tests and for the development
of automatic text simplification tools. A sample of 21 poor-reading
and dyslexic children with an average reading delay of 2.5 years
read a portion of the corpus. The transcripts of readings errors
were integrated into the corpus with the goal of identifying lexical
difficulty in the target population. By means of statistical
testing, we provide evidence that the manual simplifications
significantly reduced reading errors, highlighting that the words
targeted for simplification were not only well-chosen but also
```

substituted with substantially easier alternatives. The entire corpus is available for consultation through a web interface and available on demand for research purposes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.169}
}

@InProceedings{hge:2020:LREC,
  author   = {Höge, Harald},
  title   = {Cortical Speech Databases For Deciphering the Articulatory Code},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {133--137},
  abstract = {The paper relates to following ‘AC-hypotheses’: The articulatory code (AC) is a neural code exchanging multi-item messages between the short-term memory and cortical areas as the vSMC and STG. In these areas already neurons active in the presence of articulatory features have been measured. The AC codes the content of speech segmented in chunks and is the same for both modalities - speech perception and speech production. Each AC-message is related to a syllable. The items of each message relate to coordinated articulatory gestures composing the syllable. The mechanism to transport the AC and to segment the auditory signal is based on θ/γ-oscillations, where a θ-cycle has the duration of a θ-syllable. The paper describes the findings from neuroscience, phonetics and the science of evolution leading to the AC-hypotheses. The paper proposes to verify the AC-hypotheses by measuring the activity of all ensembles of neurons coding and decoding the AC. Due to state of the art, the cortical measurements to be prepared, done and further processed need a high effort from scientists active in different areas. We propose to launch a project to produce cortical speech databases with cortical recordings synchronized with the speech signal allowing to decipher the articulatory code.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.17}
}

@InProceedings{moseley-EtAl:2020:LREC,
  author   = {Moseley, Edward T. and Wu, Joy T. and Welt, Jonathan and Foote, John and Tyler, Patrick D. and Grant, David W. and Carlson, Eric T. and Gehrmann, Sebastian and Dernoncourt, Franck and Celi, Leo Anthony},
  title   = {A Corpus for Detecting High-Context Medical Conditions in Intensive Care Patient Notes Focusing on Frequently Readmitted Patients},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {1362--1367},
  abstract  = {A crucial step within secondary analysis of
electronic health records (EHRs) is to identify the patient cohort
under investigation. While EHRs contain medical billing codes that
aim to represent the conditions and treatments patients may have,
much of the information is only present in the patient notes.
Therefore, it is critical to develop robust algorithms to infer
patients' conditions and treatments from their written notes. In
this paper, we introduce a dataset for patient phenotyping, a task
that is defined as the identification of whether a patient has a
given medical condition (also referred to as clinical indication or
phenotype) based on their patient note. Nursing Progress Notes and
Discharge Summaries from the Intensive Care Unit of a large tertiary
care hospital were manually annotated for the presence of several
high-context phenotypes relevant to treatment and risk of re-
hospitalization. This dataset contains 1102 Discharge Summaries and
1000 Nursing Progress Notes. Each Discharge Summary and Progress
Note has been annotated by at least two expert human annotators (one
clinical researcher and one resident physician). Annotated
phenotypes include treatment non-adherence, chronic pain, advanced/
metastatic cancer, as well as 10 other phenotypes. This dataset can
be utilized for academic and industrial research in medicine and
computer science, particularly within the field of medical natural
language processing.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.170}
}

@InProceedings{zotova-EtAl:2020:LREC,
  author    = {Zotova, Elena  and  Agerri, Rodrigo  and  Nuñez,
Manuel  and  Rigau, German},
  title     = {Multilingual Stance Detection in Tweets: The
Catalonia Independence Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1368--1375},
  abstract  = {Stance detection aims to determine the attitude of a
given text with respect to a specific topic or claim. While stance
detection has been fairly well researched in the last years, most
the work has been focused on English. This is mainly due to the
relative lack of annotated data in other languages. The TW-10
referendum Dataset released at IberEval 2018 is a previous effort to
provide multilingual stance-annotated data in Catalan and Spanish.
Unfortunately, the TW-10 Catalan subset is extremely imbalanced.
This paper addresses these issues by presenting a new multilingual
dataset for stance detection in Twitter for the Catalan and Spanish
languages, with the aim of facilitating research on stance detection
in multilingual and cross-lingual settings. The dataset is annotated
with stance towards one topic, namely, the ndependence of Catalonia.
We also provide a semi-automatic method to annotate the dataset
based on a categorization of Twitter users. We experiment on the new
```

```
corpus with a number of supervised approaches, including linear
classifiers and deep learning methods. Comparison of our new corpus
with the with the TW-10 dataset shows both the benefits and
potential of a well balanced corpus for multilingual and cross-
lingual research on stance detection. Finally, we establish new
state-of-the-art results on the TW-10 dataset, both for Catalan and
Spanish.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.171}
}

@InProceedings{moeed-EtAl:2020:LREC1,
  author    = {Moeed, Abdul  and  Hagerer, Gerhard  and  Dugar,
Sumit  and  Gupta, Sarthak  and  Ghosh, Mainak  and  Danner, Hannah
and  Mitevski, Oliver  and  Nawroth, Andreas  and  Groh, Georg},
  title     = {An Evaluation of Progressive Neural Networksfor
Transfer Learning in Natural Language Processing},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1376--1381},
  abstract  = {A major challenge in modern neural networks is the
utilization of previous knowledge for new tasks in an effective
manner, otherwise known as transfer learning. Fine-tuning, the most
widely used method for achieving this, suffers from catastrophic
forgetting. The problem is often exacerbated in natural language
processing (NLP). In this work, we assess progressive neural
networks (PNNs) as an alternative to fine-tuning. The evaluation is
based on common NLP tasks such as sequence labeling and text
classification. By gauging PNNs across a range of architectures,
datasets, and tasks, we observe improvements over the baselines
throughout all experiments.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.172}
}

@InProceedings{ccillon-EtAl:2020:LREC,
  author    = {Cécillon, Noé  and  Labatut, Vincent  and  Dufour,
Richard  and  Linarès, Georges},
  title     = {WAC: A Corpus of Wikipedia Conversations for Online
Abuse Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1382--1390},
  abstract  = {With the spread of online social networks, it is more
and more difficult to monitor all the user-generated content.
Automating the moderation process of the inappropriate exchange
content on Internet has thus become a priority task. Methods have
been proposed for this purpose, but it can be challenging to find a
```

suitable dataset to train and develop them. This issue is especially true for approaches based on information derived from the structure and the dynamic of the conversation. In this work, we propose an original framework, based on the the Wikipedia Comment corpus, with comment-level abuse annotations of different types. The major contribution concerns the reconstruction of conversations, by comparison to existing corpora, which focus only on isolated messages (i.e. taken out of their conversational context). This large corpus of more than 380k annotated messages opens perspectives for online abuse detection and especially for context-based approaches. We also propose, in addition to this corpus, a complete benchmarking platform to stimulate and fairly compare scientific works around the problem of content abuse detection, trying to avoid the recurring problem of result replication. Finally, we apply two classification methods to our dataset to demonstrate its potential.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.173}
}

@InProceedings{hamoui-mars-almotairi:2020:LREC,
  author   = {Hamoui, Btool and Mars, Mourad and Almotairi, Khaled},
  title   = {FloDusTA: Saudi Tweets Dataset for Flood, Dust Storm, and Traffic Accident Events},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1391--1396},
  abstract = {The rise of social media platforms makes it a valuable information source of recent events and users' perspective towards them. Twitter has been one of the most important communication platforms in recent years. Event detection, one of the information extraction aspects, involves identifying specified types of events in the text. Detecting events from tweets can help to predict real-world events precisely. A serious challenge that faces Arabic event detection is the lack of Arabic datasets that can be exploited in detecting events. This paper will describe FloDusTA, which is a dataset of tweets that we have built for the purpose of developing an event detection system. The dataset contains tweets written in both Modern Standard Arabic and Saudi dialect. The process of building the dataset starting from tweets collection to annotation by human annotators will be present. The tweets are labeled with four labels: flood, dust storm, traffic accident, and non-event. The dataset was tested for classification and the result was strongly encouraging.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.174}
}

@InProceedings{chiril-EtAl:2020:LREC,
  author   = {Chiril, Patricia and Moriceau, Véronique and Benamara, Farah and Mari, Alda and Origgi, Gloria and Coulomb-

Gully, Marlène},
  title   = {An Annotated Corpus for Sexism Detection in French Tweets},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1397--1403},
  abstract = {Social media networks have become a space where users are free to relate their opinions and sentiments which may lead to a large spreading of hatred or abusive messages which have to be moderated. This paper presents the first French corpus annotated for sexism detection composed of about 12,000 tweets. In a context of offensive content mediation on social media now regulated by European laws, we think that it is important to be able to detect automatically not only sexist content but also to identify if a message with a sexist content is really sexist (i.e. addressed to a woman or describing a woman or women in general) or is a story of sexism experienced by a woman. This point is the novelty of our annotation scheme. We also propose some preliminary results for sexism detection obtained with a deep learning approach. Our experiments show encouraging results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.175}
}

@InProceedings{santos-EtAl:2020:LREC,
  author   = {Santos, Roney and Pedro, Gabriela and Leal, Sidney and Vale, Oto and Pardo, Thiago and Bontcheva, Kalina and Scarton, Carolina},
  title   = {Measuring the Impact of Readability Features in Fake News Detection},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1404--1413},
  abstract = {The proliferation of fake news is a current issue that influences a number of important areas of society, such as politics, economy and health. In the Natural Language Processing area, recent initiatives tried to detect fake news in different ways, ranging from language-based approaches to content-based verification. In such approaches, the choice of the features for the classification of fake and true news is one of the most important parts of the process. This paper presents a study on the impact of readability features to detect fake news for the Brazilian Portuguese language. The results show that such features are relevant to the task (achieving, alone, up to 92\% classification accuracy) and may improve previous classification results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.176}
}

```
@InProceedings{stajner-hulpu:2020:LREC,
  author    = {Stajner, Sanja  and  Hulpuș, Ioana},
  title     = {When Shallow is Good Enough: Automatic Assessment of
Conceptual Text Complexity using Shallow Semantic Features},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1414--1422},
  abstract  = {According to psycholinguistic studies, the complexity
of concepts used in a text and the relations between mentioned
concepts play the most important role in text understanding and
maintaining reader's interest. However, the classical approaches to
automatic assessment of text complexity, and their commercial
applications, take into consideration mainly syntactic and lexical
complexity. Recently, we introduced the task of automatic assessment
of conceptual text complexity, proposing a set of graph-based deep
semantic features using DBpedia as a proxy to human knowledge. Given
that such graphs can be noisy, incomplete, and computationally
expensive to deal with, in this paper, we propose the use of textual
features and shallow semantic features that only require entity
linking. We compare the results obtained with new features with
those of the state-of-the-art deep semantic features on two tasks:
(1) pairwise comparison of two versions of the same text; and (2)
five-level classification of texts. We find that the shallow
features achieve state-of-the-art results on both tasks,
significantly outperforming performances of the deep semantic
features on the five-level classification task. Interestingly, the
combination of the shallow and deep semantic features lead to a
significant improvement of the performances on that task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.177}
}

@InProceedings{capuozzo-EtAl:2020:LREC,
  author    = {Capuozzo, Pasquale  and  Lauriola, Ivano  and
Strapparava, Carlo  and  Aiolli, Fabio  and  Sartori, Giuseppe},
  title     = {DecOp: A Multilingual and Multi-domain Corpus For
Detecting Deception In Typed Text},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1423--1430},
  abstract  = {In recent years, the increasing interest in the
development of automatic approaches for unmasking deception in
online sources led to promising results. Nonetheless, among the
others, two major issues remain still unsolved: the stability of
classifiers performances across different domains and languages.
Tackling these issues is challenging since labelled corpora
```

involving multiple domains and compiled in more than one language are few in the scientific literature. For filling this gap, in this paper we introduce DecOp (Deceptive Opinions), a new language resource developed for automatic deception detection in cross-domain and cross-language scenarios. DecOp is composed of 5000 examples of both truthful and deceitful first-person opinions balanced both across five different domains and two languages and, to the best of our knowledge, is the largest corpus allowing cross-domain and cross-language comparisons in deceit detection tasks. In this paper, we describe the collection procedure of the DecOp corpus and his main characteristics. Moreover, the human performance on the DecOp test-set and preliminary experiments by means of machine learning models based on Transformer architecture are shown.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.178}
}

@InProceedings{blandin-EtAl:2020:LREC,
  author   = {Blandin, Alexis and Lecorvé, Gwénolé and Battistelli, Delphine and Étienne, Aline},
  title   = {Age Recommendation for Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1431--1439},
  abstract = {The understanding of a text by a reader or listener is conditioned by the adequacy of the text's characteristics with the person's capacities and knowledge. This adequacy is critical in the case of a child since her/his cognitive and linguistic skills are still under development. Hence, in this paper, we present and study an original natural language processing (NLP) task which consists in predicting the age from which a text can be understood by someone. To do so, this paper first exhibits features derived from the psycholinguistic domain, as well as some coming from related NLP tasks. Then, we propose a set of neural network models and compare them on a dataset of French texts dedicated to young or adult audiences. To circumvent the lack of data, we study the idea to predict ages at the sentence level. The experiments first show that the sentence-based age recommendations can be efficiently merged to predict text-based recommendations. Then, we also demonstrate that the age predictions returned by our best model are better than those provided by psycholinguists. Finally, the paper investigates the impact of the various features used in these results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.179}
}

@InProceedings{hollenstein-EtAl:2020:LREC,
  author   = {Hollenstein, Nora and Troendle, Marius and Zhang, Ce and Langer, Nicolas},
  title   = {ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {138--146},
  abstract  = {We recorded and preprocessed ZuCo 2.0, a new dataset
of simultaneous eye-tracking and electroencephalography during
natural reading and during annotation. This corpus contains gaze and
brain activity data of 739 English sentences, 349 in a normal
reading paradigm and 390 in a task-specific paradigm, in which the
18 participants actively search for a semantic relation type in the
given sentences as a linguistic annotation task. This new dataset
complements ZuCo 1.0 by providing experiments designed to analyze
the differences in cognitive processing between natural reading and
annotation. The data is freely available here: https://osf.io/
2urht/.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.18}
}

@InProceedings{huang-EtAl:2020:LREC1,
  author    = {Huang, Xiaolei  and  Xing, Linzi  and  Dernoncourt,
Franck  and  Paul, Michael J.},
  title     = {Multilingual Twitter Corpus and Baselines for
Evaluating Demographic Bias in Hate Speech Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1440--1448},
  abstract  = {Existing research on fairness evaluation of document
classification models mainly uses synthetic monolingual data without
ground truth for author demographic attributes. In this work, we
assemble and publish a multilingual Twitter corpus for the task of
hate speech detection with inferred four author demographic factors:
age, country, gender and race/ethnicity. The corpus covers five
languages: English, Italian, Polish, Portuguese and Spanish. We
evaluate the inferred demographic labels with a crowdsourcing
platform, Figure Eight. To examine factors that can cause biases, we
take an empirical analysis of demographic predictability on the
English corpus. We measure the performance of four popular document
classifiers and evaluate the fairness and bias of the baseline
classifiers on the author-level demographic attributes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.180}
}

@InProceedings{luzdearaujo-EtAl:2020:LREC,
  author    = {Luz de Araujo, Pedro Henrique  and  de Campos,
Teófilo Emídio  and  Ataides Braz, Fabricio  and  Correia da Silva,
Nilton},
  title     = {VICTOR: a Dataset for Brazilian Legal Documents
```

Classification},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1449--1458},
  abstract = {This paper describes VICTOR, a novel dataset built from Brazil's Supreme Court digitalized legal documents, composed of more than 45 thousand appeals, which includes roughly 692 thousand documents---about 4.6 million pages. The dataset contains labeled text data and supports two types of tasks: document type classification; and theme assignment, a multilabel problem. We present baseline results using bag-of-words models, convolutional neural networks, recurrent neural networks and boosting algorithms. We also experiment using linear-chain Conditional Random Fields to leverage the sequential nature of the lawsuits, which we find to lead to improvements on document type classification. Finally we compare a theme classification approach where we use domain knowledge to filter out the less informative document pages to the default one where we use all pages. Contrary to the Court experts' expectations, we find that using all available data is the better method. We make the dataset available in three versions of different sizes and contents to encourage explorations of better models and techniques.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.181}
}

@InProceedings{patel-caragea-phillips:2020:LREC,
  author   = {Patel, Krutarth and Caragea, Cornelia and Phillips, Mark},
  title   = {Dynamic Classification in Web Archiving Collections},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1459--1468},
  abstract = {The Web archived data usually contains high-quality documents that are very useful for creating specialized collections of documents. To create such collections, there is a substantial need for automatic approaches that can distinguish the documents of interest for a collection out of the large collections (of millions in size) from Web Archiving institutions. However, the patterns of the documents of interest can differ substantially from one document to another, which makes the automatic classification task very challenging. In this paper, we explore dynamic fusion models to find, on the fly, the model or combination of models that performs best on a variety of document types. Our experimental results show that the approach that fuses different models outperforms individual models and other ensemble methods on three datasets.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.182}

}

@InProceedings{vasconcelos-campelo-jeronimo:2020:LREC,
  author   = {Vasconcelos, Larissa and Campelo, Claudio and Jeronimo, Caio},
  title   = {Aspect Flow Representation and Audio Inspired Analysis for Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1469--1477},
  abstract = {For better understanding how people write texts, it is fundamental to examine how a particular aspect (e.g., subjectivity, sentiment, argumentation) is exploited in a text. Analysing such an aspect of a text as a whole (i.e., through a summarised single feature) can lead to significant information loss. In this paper, we propose a novel method of representing and analysing texts that consider how an aspect behaves throughout the text. We represent the texts by aspect flows for capturing all the aspect behaviour. Then, inspired by the resemblance between these flows format and a sound waveform, we fragment them into frames and calculate an adaptation of audio analysis features, named here Audio-Like Features, as a way of analysing the texts. The results of the conducted classification tasks reveal that our approach can surpass methods based on summarised features. We also show that a detailed examination of the Audio-Like Features can lead to a more profound knowledge about the represented texts.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.183}
}

@InProceedings{lim-EtAl:2020:LREC,
  author   = {Lim, Sora and Jatowt, Adam and Färber, Michael and Yoshikawa, Masatoshi},
  title   = {Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1478--1484},
  abstract = {The spread of biased news and its consumption by the readers has become a considerable issue. Researchers from multiple domains including social science and media studies have made efforts to mitigate this media bias issue. Specifically, various techniques ranging from natural language processing to machine learning have been used to help determine news bias automatically. However, due to the lack of publicly available datasets in this field, especially ones containing labels concerning bias on a fine-grained level (e.g., on sentence level), it is still challenging to develop

methods for effectively identifying bias embedded in new articles. In this paper, we propose a novel news bias dataset which facilitates the development and evaluation of approaches for detecting subtle bias in news articles and for understanding the characteristics of biased sentences. Our dataset consists of 966 sentences from 46 English-language news articles covering 4 different events and contains labels concerning bias on the sentence level. For scalability reasons, the labels were obtained based on crowd-sourcing. Our dataset can be used for analyzing news bias, as well as for developing and evaluating methods for news bias detection. It can also serve as resource for related researches including ones focusing on fake news detection.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.184}
}

@InProceedings{jayashree-srijith:2020:LREC,
  author   = {Jayashree, P. and Srijith, P. K.},
  title   = {Evaluation of Deep Gaussian Processes for Text Classification},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1485--1491},
  abstract = {With the tremendous success of deep learning models on computer vision tasks, there are various emerging works on the Natural Language Processing (NLP) task of Text Classification using parametric models. However, it constrains the expressability limit of the function and demands enormous empirical efforts to come up with a robust model architecture. Also, the huge parameters involved in the model causes over-fitting when dealing with small datasets. Deep Gaussian Processes (DGP) offer a Bayesian non-parametric modelling framework with strong function compositionality, and helps in overcoming these limitations. In this paper, we propose DGP models for the task of Text Classification and an empirical comparison of the performance of shallow and Deep Gaussian Process models is made. Extensive experimentation is performed on the benchmark Text Classification datasets such as TREC (Text REtrieval Conference), SST (Stanford Sentiment Treebank), MR (Movie Reviews), R8 (Reuters-8), which demonstrate the effectiveness of DGP models.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.185}
}

@InProceedings{plazadelarco-EtAl:2020:LREC,
  author   = {Plaza del Arco, Flor Miriam and Strapparava, Carlo and Urena Lopez, L. Alfonso and Martin, Maite},
  title   = {EmoEvent: A Multilingual Emotion Corpus based on different Events},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1492--1498},
  abstract  = {In recent years emotion detection in text has become
more popular due to its potential applications in fields such as
psychology, marketing, political science, and artificial
intelligence, among others. While opinion mining is a well-
established task with many standard data sets and well-defined
methodologies, emotion mining has received less attention due to its
complexity. In particular, the annotated gold standard resources
available are not enough. In order to address this shortage, we
present a multilingual emotion data set based on different events
that took place in April 2019. We collected tweets from the Twitter
platform. Then one of seven emotions, six Ekman's basic emotions
plus the ``neutral or other emotions", was labeled on each tweet by
3 Amazon MTurkers. A total of 8,409 in Spanish and 7,303 in English
were labeled. In addition, each tweet was also labeled as offensive
or no offensive. We report some linguistic statistics about the data
set in order to observe the difference between English and Spanish
speakers when they express emotions related to the same events.
Moreover, in order to validate the effectiveness of the data set, we
also propose a machine learning approach for automatically detecting
emotions in tweets for both languages, English and Spanish.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.186}
}

@InProceedings{jaiswal-EtAl:2020:LREC,
  author    = {Jaiswal, Mimansa  and  Bara, Cristian-Paul  and  Luo,
Yuanhang  and  Burzo, Mihai  and  Mihalcea, Rada  and  Provost,
Emily Mower},
  title     = {MuSE: a Multimodal Dataset of Stressed Emotion},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1499--1510},
  abstract  = {Endowing automated agents with the ability to provide
support, entertainment and interaction with human beings requires
sensing of the users' affective state. These affective states are
impacted by a combination of emotion inducers, current psychological
state, and various conversational factors. Although emotion
classification in both singular and dyadic settings is an
established area, the effects of these additional factors on the
production and perception of emotion is understudied. This paper
presents a new dataset, Multimodal Stressed Emotion (MuSE), to study
the multimodal interplay between the presence of stress and
expressions of affect. We describe the data collection protocol, the
possible areas of use, and the annotations for the emotional content
of the recordings. The paper also presents several baselines to
measure the performance of multimodal features for emotion and
stress classification.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.187}
```

```
}

@InProceedings{zhang-EtAl:2020:LREC1,
  author    = {Zhang, Linrui  and  Huang, Hsin-Lun  and  Yu, Yang
and  Moldovan, Dan},
  title     = {Affect inTweets: A Transfer Learning Approach},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1511--1516},
  abstract  = {People convey sentiments and emotions through
language. To understand these affectual states is an essential step
towards understanding natural language. In this paper, we propose a
transfer-learning based approach to inferring the affectual state of
a person from their tweets. As opposed to the traditional machine
learning models which require considerable effort in designing task
specific features, our model can be well adapted to the proposed
tasks with a very limited amount of fine-tuning, which significantly
reduces the manual effort in feature engineering. We aim to show
that by leveraging the pre-learned knowledge, transfer learning
models can achieve competitive results in the affectual content
analysis of tweets, compared to the traditional models. As shown by
the experiments on SemEval-2018 Task 1: Affect in Tweets, our model
ranking 2nd, 4th and 6th place in four of its subtasks proves the
effectiveness of our idea.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.188}
}

@InProceedings{tammewar-EtAl:2020:LREC,
  author    = {Tammewar, Aniruddha  and  Cervone, Alessandra  and
Messner, Eva-Maria  and  Riccardi, Giuseppe},
  title     = {Annotation of Emotion Carriers in Personal
Narratives},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1517--1525},
  abstract  = {We are interested in the problem of understanding
personal narratives (PN) - spoken or written - recollections of
facts, events, and thoughts. For PNs, we define emotion carriers as
the speech or text segments that best explain the emotional state of
the narrator. Such segments may span from single to multiple words,
containing for example verb or noun phrases. Advanced automatic
understanding of PNs requires not only the prediction of the
narrator's emotional state but also to identify which events (e.g.
the loss of a relative or the visit of grandpa) or people (e.g. the
old group of high school mates) carry the emotion manifested during
the personal recollection. This work proposes and evaluates an
```

annotation model for identifying emotion carriers in spoken personal narratives. Compared to other text genres such as news and microblogs, spoken PNs are particularly challenging because a narrative is usually unstructured, involving multiple sub-events and characters as well as thoughts and associated emotions perceived by the narrator. In this work, we experiment with annotating emotion carriers in speech transcriptions from the Ulm State-of-Mind in Speech (USoMS) corpus, a dataset of PNs in German. We believe this resource could be used for experiments in the automatic extraction of emotion carriers from PN, a task that could provide further advancements in narrative understanding.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.189}
}

@InProceedings{reinboth-EtAl:2020:LREC,
  author   = {Reinboth, Tim and Gross, Stephanie and Bishop, Laura and Krenn, Brigitte},
  title   = {Linguistic, Kinematic and Gaze Information in Task Descriptions: The LKG-Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {147--155},
  abstract = {Data from neuroscience and psychology suggest that sensorimotor cognition may be of central importance to language. Specifically, the linguistic structure of utterances referring to concrete actions may reflect the structure of the sensorimotor processing underlying the same action. To investigate this, we present the Linguistic, Kinematic and Gaze information in task descriptions Corpus (LKG-Corpus), comprising multimodal data on 13 humans, conducting take, put, and push actions, and describing these actions with 350 utterances. Recorded are audio, video, motion and eye-tracking data while participants perform an action and describe what they do. The dataset is annotated with orthographic transcriptions of utterances and information on: (a) gaze behaviours, (b) when a participant touched an object, (c) when an object was moved, (d) when a participant looked at the location s/he would next move the object to, (e) when the participant’s gaze was stable on an area. With the exception of the annotation of stable gaze, all annotations were performed manually. With the LKG-Corpus, we present a dataset that integrates linguistic, kinematic and gaze data with an explicit focus on relations between action and language. On this basis, we outline applications of the dataset to both basic and applied research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.19}
}

@InProceedings{wottawa-EtAl:2020:LREC,
  author   = {Wottawa, Jane and Tahon, Marie and Marin, Apolline and Audibert, Nicolas},
  title   = {Towards Interactive Annotation for Hesitation in

Conversational Speech},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1526--1532},
  abstract = {Manual annotation of speech corpora is expensive in both human resources and time. Furthermore, recognizing affects in spontaneous, non acted speech presents a challenge for humans and machines. The aim of the present study is to automatize the labeling of hesitant speech as a marker of expressed uncertainty. That is why, the NCCFr-corpus was manually annotated for 'degree of hesitation' on a continuous scale between -3 and 3 and the affective dimensions 'activation, valence and control'. In total, 5834 chunks of the NCCFr-corpus were manually annotated. Acoustic analyses were carried out based on these annotations. Furthermore, regression models were trained in order to allow automatic prediction of hesitation for speech chunks that do not have a manual annotation. Preliminary results show that the number of filled pauses as well as vowel duration increase with the degree of hesitation, and that automatic prediction of the hesitation degree reaches encouraging RMSE results of 1.6.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.190}
}

@InProceedings{costajuss-EtAl:2020:LREC,
  author   = {Costa-jussà, Marta R. and González, Esther and Moreno, Asuncion and Cumalat, Eudald},
  title   = {Abusive language in Spanish children and young teenager’s conversations: data preparation and short text classification with contextual word embeddings},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1533--1537},
  abstract = {Abusive texts are reaching the interests of the scientific and social community. How to automatically detect them is onequestion that is gaining interest in the natural language processing community. The main contribution of this paper is toevaluate the quality of the recently developed ”Spanish Database for cyberbullying prevention” for the purpose of trainingclassifiers on detecting abusive short texts. We compare classical machine learning techniques to the use of a more ad-vanced model: the contextual word embeddings in the particular case of classification of abusive short-texts for the Spanishlanguage. As contextual word embeddings, we use Bidirectional Encoder Representation from Transformers (BERT), pro-posed at the end of 2018. We show that BERT mostly outperforms classical techniques. Far beyond the experimentalimpact of our research, this project aims at planting

the seeds for an innovative technological tool with a high potentialsocial impact and aiming at being part of the initiatives in artificial intelligence for social good.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.191}
}

@InProceedings{rambabu-EtAl:2020:LREC,
  author   = {Rambabu, Banothu and Botsa, Kishore Kumar and Paidi, Gangamohan and Gangashetty, Suryakanth V},
  title   = {IIIT-H TEMD Semi-Natural Emotional Speech Database from Professional Actors and Non-Actors},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1538--1545},
  abstract = {A fundamental essence for emotional speech analysis towards emotion recognition is a good database. Database collected from natural scenarios consists of spontaneous emotions, but there are several issues in collection of such database. Other than the privacy and legal related concerns, there is no control over environment at the background. As it is difficult to collect data from natural scenarios, many research groups have collected data from semi-natural or designed procedures. In this paper, a new emotional speech database named IIIT-H TEMD (International Institute of Information Technology-Hyderabad Telugu Emotional Database) is collected using designed drama situations from actors and non-actors. Utterances are manually annotated using a hybrid strategy by giving the context to one of the listeners. As some of the data collection studies in the literature recommend for actors, analysis of actors data versus non-actors data is carried out for their significance. The total size of the dataset is about 5 hours, which makes it an useful resource for the emotional speech analysis.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.192}
}

@InProceedings{janssoone-EtAl:2020:LREC,
  author   = {Janssoone, Thomas and Bailly, Kévin and Richard, Gaël and Clavel, Chloé},
  title   = {The POTUS Corpus, a Database of Weekly Addresses for the Study of Stance in Politics and Virtual Agents},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1546--1553},
  abstract = {One of the main challenges in the field of Embodied Conversational Agent (ECA) is to generate socially believable agents. The common strategy for agent behaviour synthesis is to rely on dedicated corpus analysis. Such a corpus is composed of

multimedia files of socio-emotional behaviors which have been annotated by external observers. The underlying idea is to identify interaction information for the agent’s socio-emotional behavior by checking whether the intended socio-emotional behavior is actually perceived by humans. Then, the annotations can be used as learning classes for machine learning algorithms applied to the social signals. This paper introduces the POTUS Corpus composed of high-quality audio-video files of political addresses to the American people. Two protagonists are present in this database. First, it includes speeches of former president Barack Obama to the American people. Secondly, it provides videos of these same speeches given by a virtual agent named Rodrigue. The ECA reproduces the original address as closely as possible using social signals automatically extracted from the original one. Both are annotated for social attitudes, providing information about the stance observed in each file. It also provides the social signals automatically extracted from Obama’s addresses used to generate Rodrigue’s ones.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.193}
}

@InProceedings{bostan-kim-klinger:2020:LREC,
  author   = {Bostan, Laura Ana Maria and Kim, Evgeny and Klinger, Roman},
  title   = {GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1554--1566},
  abstract = {Most research on emotion analysis from text focuses on the task of emotion classification or emotion intensity regression. Fewer works address emotions as a phenomenon to be tackled with structured learning, which can be explained by the lack of relevant datasets. We fill this gap by releasing a dataset of 5000 English news headlines annotated via crowdsourcing with their associated emotions, the corresponding emotion experiencers and textual cues, related emotion causes and targets, as well as the reader’s perception of the emotion of the headline. This annotation task is comparably challenging, given the large number of classes and roles to be identified. We therefore propose a multiphase annotation procedure in which we first find relevant instances with emotional content and then annotate the more fine-grained aspects. Finally, we develop a baseline for the task of automatic prediction of semantic role structures and discuss the results. The corpus we release enables further research on emotion classification, emotion intensity prediction, emotion cause detection, and supports further qualitative studies.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.194}
}

@InProceedings{kiritchenko-EtAl:2020:LREC,

```
  author    = {Kiritchenko, Svetlana  and  Hipson, Will  and
Coplan, Robert  and  Mohammad, Saif M.},
  title     = {SOLO: A Corpus of Tweets for Examining the State of
Being Alone},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1567--1577},
  abstract  = {The state of being alone can have a substantial
impact on our lives, though experiences with time alone diverge
significantly among individuals. Psychologists distinguish between
the concept of solitude, a positive state of voluntary aloneness,
and the concept of loneliness, a negative state of dissatisfaction
with the quality of one's social interactions. Here, for the first
time, we conduct a large-scale computational analysis to explore how
the terms associated with the state of being alone are used in
online language. We present SOLO (State of Being Alone), a corpus of
over 4 million tweets collected with query terms solitude, lonely,
and loneliness. We use SOLO to analyze the language and emotions
associated with the state of being alone. We show that the term
solitude tends to co-occur with more positive, high-dominance words
(e.g., enjoy, bliss) while the terms lonely and loneliness
frequently co-occur with negative, low-dominance words (e.g.,
scared, depressed), which confirms the conceptual distinctions made
in psychology. We also show that women are more likely to report on
negative feelings of being lonely as compared to men, and there are
more teenagers among the tweeters that use the word lonely than
among the tweeters that use the word solitude.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.195}
}

@InProceedings{hipson-mohammad:2020:LREC,
  author    = {Hipson, Will  and  Mohammad, Saif M.},
  title     = {PoKi: A Large Dataset of Poems by Children},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1578--1589},
  abstract  = {Child language studies are crucial in improving our
understanding of child well-being; especially in determining the
factors that impact happiness, the sources of anxiety, techniques of
emotion regulation, and the mechanisms to cope with stress. However,
much of this research is stymied by the lack of availability of
large child-written texts. We present a new corpus of child-written
text, PoKi, which includes about 62 thousand poems written by
children from grades 1 to 12. PoKi is especially useful in studying
child language because it comes with information about the age of
the child authors (their grade). We analyze the words in PoKi along
```

```
several emotion dimensions (valence, arousal, dominance) and
discrete emotions (anger, fear, sadness, joy). We use non-parametric
regressions to model developmental differences from early childhood
to late-adolescence. Results show decreases in valence that are
especially pronounced during mid-adolescence, while arousal and
dominance peaked during adolescence. Gender differences in the
developmental trajectory of emotions are also observed. Our results
support and extend the current state of emotion development
research.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.196}
}

@InProceedings{macary-EtAl:2020:LREC,
  author    = {Macary, Manon  and  Tahon, Marie  and  Estève,
Yannick  and  Rousseau, Anthony},
  title     = {AlloSat: A New Call Center French Corpus for
Satisfaction and Frustration Analysis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1590--1597},
  abstract  = {We present a new corpus, named AlloSat, composed of
real-life call center conversations in French that is continuously
annotated in frustration and satisfaction. This corpus has been set
up to develop new systems able to model the continuous aspect of
semantic and paralinguistic information at the conversation level.
The present work focuses on the paralinguistic level, more precisely
on the expression of emotions. In the call center industry, the
conversation usually aims at solving the caller's request. As far as
we know, most emotional databases contain static annotations in
discrete categories or in dimensions such as activation or valence.
We hypothesize that these dimensions are not task-related enough.
Moreover, static annotations do not enable to explore the temporal
evolution of emotional states. To solve this issue, we propose a
corpus with a rich annotation scheme enabling a real-time
investigation of the axis frustration / satisfaction. AlloSat
regroups 303 conversations with a total of approximately 37 hours of
audio, all recorded in real-life environments collected by Allo-
Media (an intelligent call tracking company). First regression
experiments, with audio features, show that the evolution of
frustration / satisfaction axis can be retrieved automatically at
the conversation level.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.197}
}

@InProceedings{wu-chien:2020:LREC,
  author    = {Wu, Shih-Hung  and  Chien, Sheng-Lun},
  title     = {Learning the Human Judgment for the Automatic
Evaluation of Chatbot},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1598--1602},
  abstract  = {It is hard to evaluate the quality of the generated
text by a generative dialogue system. Currently, dialogue evaluation
relies on human judges to label the quality of the generated text.
It is not a reusable mechanism that can give consistent evaluation
for system developers. We believe that it is easier to get
consistent results on comparing two generated dialogue by two
systems and it is hard to give a consistent quality score on only
one system at a time. In this paper, we propose a machine learning
approach to reduce the effort of human evaluation by learning the
human judgment on comparing two dialogue systems. Training from the
human labeling result, the evaluation model learns which generative
models is better in each dialog context. Thus, it can be used for
system developers to compare the fine-tuned models over and over
again without the human labor. In our experiment we find the
agreement between the learned model and human judge is 70\%. The
experiment is conducted on comparing two attention based GRU-RNN
generative models.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.198}
}

@InProceedings{lee-lim-choi:2020:LREC,
  author    = {Lee, Young-Jun  and  Lim, Chae-Gyun  and  Choi, Ho-
Jin},
  title     = {Korean-Specific Emotion Annotation Procedure Using N-
Gram-Based Distant Supervision and Korean-Specific-Feature-Based
Distant Supervision},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1603--1610},
  abstract  = {Detecting emotions from texts is considerably
important in an NLP task, but it has the limitation of the scarcity
of manually labeled data. To overcome this limitation, many
researchers have annotated unlabeled data with certain frequently
used annotation procedures. However, most of these studies are
focused mainly on English and do not consider the characteristics of
the Korean language. In this paper, we present a Korean-specific
annotation procedure, which consists of two parts, namely n-gram-
based distant supervision and Korean-specific-feature-based distant
supervision. We leverage the distant supervision with the n-gram and
Korean emotion lexicons. Then, we consider the Korean-specific
emotion features. Through experiments, we showed the effectiveness
of our procedure by comparing with the KTEA dataset. Additionally,
we constructed a large-scale emotion-labeled dataset, Korean Movie
Review Emotion (KMRE) Dataset, using our procedure. In order to
construct our dataset, we used a large-scale sentiment movie review
```

```
corpus as the unlabeled dataset. Moreover, we used a Korean emotion
lexicon provided by KTEA. We also performed an emotion
classification task and a human evaluation on the KMRE dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.199}
}

@InProceedings{yu-uma-poesio:2020:LREC,
  author    = {Yu, Juntao  and  Uma, Alexandra  and  Poesio,
Massimo},
  title     = {A Cluster Ranking Model for Full Anaphora
Resolution},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {11--20},
  abstract  = {Anaphora resolution (coreference) systems designed
for the CONLL 2012 dataset typically cannot handle key aspects of
the full anaphora resolution task such as the identification of
singletons and of certain types of non-referring expressions (e.g.,
expletives), as these aspects are not annotated in that corpus.
However, the recently released dataset for the CRAC 2018 Shared Task
can now be used for that purpose. In this paper, we introduce an
architecture to simultaneously identify non-referring expressions
(including expletives, predicative {\NP}s, and other types) and
build coreference chains, including singletons. Our cluster-ranking
system uses an attention mechanism to determine the relative
importance of the mentions in the same cluster. Additional
classifiers are used to identify singletons and non-referring
markables. Our contributions are as follows. First all, we report
the first result on the CRAC data using system mentions; our result
is 5.8\% better than the shared task baseline system, which used
gold mentions. Second, we demonstrate that the availability of
singleton clusters and non-referring expressions can lead to
substantially improved performance on non-singleton clusters as
well. Third, we show that despite our model not being designed
specifically for the CONLL data, it achieves a score equivalent to
that of the state-of-the-art system by Kantor and Globerson (2019)
on that dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.2}
}

@InProceedings{jancso-moran-stoll:2020:LREC,
  author    = {Jancso, Anna  and  Moran, Steven  and  Stoll,
Sabine},
  title     = {The ACQDIV Corpus Database and Aggregation Pipeline},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {156--165},
  abstract  = {We present the ACQDIV corpus database and aggregation
pipeline, a tool developed as part of the European Research Council
(ERC) funded project ACQDIV, which aims to identify the universal
cognitive processes that allow children to acquire any language. The
corpus database represents 15 corpora from 14 typologically
maximally diverse languages. Here we give an overview of the
project, database, and our extensible software package for adding
more corpora to the current language sample. Lastly, we discuss how
we use the corpus database to mine for universal patterns in child
language acquisition corpora and we describe avenues for future
research.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.20}
}

@InProceedings{xu-EtAl:2020:LREC1,
  author    = {Xu, Jiajun  and  Masuda, Kyosuke  and  Nishizaki,
Hiromitsu  and  Fukumoto, Fumiyo  and  Suzuki, Yoshimi},
  title     = {Semi-Automatic Construction and Refinement of an
Annotated Corpus for a Deep Learning Framework for Emotion
Classification},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1611--1617},
  abstract  = {In the case of using a deep learning (machine
learning) framework for emotion classification, one significant
difficulty faced is the requirement of building a large, emotion
corpus in which each sentence is assigned emotion labels. As a
result, there is a high cost in terms of time and money associated
with the construction of such a corpus. Therefore, this paper
proposes a method of creating a semi-automatically constructed
emotion corpus. For the purpose of this study sentences were mined
from Twitter using some emotional seed words that were selected from
a dictionary in which the emotion words were well-defined. Tweets
were retrieved by one emotional seed word, and the retrieved
sentences were assigned emotion labels based on the emotion category
of the seed word. It was evident from the findings that the deep
learning-based emotion classification model could not achieve high
levels of accuracy in emotion classification because the semi-
automatically constructed corpus had many errors when assigning
emotion labels. In this paper, therefore, an approach for improving
the quality of the emotion labels by automatically correcting the
errors of emotion labels is proposed and tested. The experimental
results showed that the proposed method worked well, and the
classification accuracy rate was improved to 55.1\% from 44.9\% on
the Twitter emotion classification task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.200}
}

@InProceedings{ghosh-ekbal-bhattacharyya:2020:LREC,
```

```
  author    = {Ghosh, Soumitra  and  Ekbal, Asif  and
Bhattacharyya, Pushpak},
  title     = {CEASE, a Corpus of Emotion Annotated Suicide notes in
English},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1618--1626},
  abstract  = {A suicide note is usually written shortly before the
suicide and it provides a chance to comprehend the self-destructive
state of mind of the deceased. From a psychological point of view,
suicide notes have been utilized for recognizing the motive behind
the suicide. To the best of our knowledge, there is no openly
accessible suicide note corpus at present, making it challenging for
the researchers and developers to deep dive into the area of mental
health assessment and suicide prevention. In this paper, we create a
fine-grained emotion annotated corpus (CEASE) of suicide notes in
English and develop various deep learning models to perform emotion
detection on the curated dataset. The corpus consists of 2393
sentences from around 205 suicide notes collected from various
sources. Each sentence is annotated with a particular emotion class
from a set of 15 fine-grained emotion labels, namely (forgiveness,
happiness\_peacefulness, love, pride, hopefulness, thankfulness,
blame, anger, fear, abuse, sorrow, hopelessness, guilt, information,
instructions). For the evaluation, we develop an ensemble
architecture, where the base models correspond to three supervised
deep learning models, namely Convolutional Neural Network (CNN),
Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM). We
obtain the highest test accuracy of 60.17\% and cross-validation
accuracy of 60.32\%},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.201}
}

@InProceedings{guhr-EtAl:2020:LREC,
  author    = {Guhr, Oliver  and  Schumann, Anne-Kathrin  and
Bahrmann, Frank  and  Böhme, Hans Joachim},
  title     = {Training a Broad-Coverage German Sentiment
Classification Model for Dialog Systems},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1627--1632},
  abstract  = {This paper describes the training of a general-
purpose German sentiment classification model. Sentiment
classification is an important aspect of general text analytics.
Furthermore, it plays a vital role in dialogue systems and voice
interfaces that depend on the ability of the system to pick up and
understand emotional signals from user utterances. The presented
```

```
study outlines how we have collected a new German sentiment corpus
and then combined this corpus with existing resources to train a
broad-coverage German sentiment model. The resulting data set
contains 5.4 million labelled samples. We have used the data to
train both, a simple convolutional and a transformer-based
classification model and compared the results achieved on various
training configurations. The model and the data set will be
published along with this paper.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.202}
}

@InProceedings{lee-lau:2020:LREC,
  author    = {Lee, Sophia Yat Mei  and  Lau, Helena Yan Ping},
  title     = {An Event-comment Social Media Corpus for Implicit
Emotion Analysis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1633--1642},
  abstract  = {The classification of implicit emotions in text has
always been a great challenge to emotion processing. Even though the
majority of emotion expressed implicitly, most previous attempts at
emotions have focused on the examination of explicit emotions. The
poor performance of existing emotion identification and
classification models can partly be attributed to the disregard of
implicit emotions. In view of this, this paper presents the
development of a Chinese event-comment social media emotion corpus.
The corpus deals with both explicit and implicit emotions with more
emphasis being placed on the implicit ones. This paper specifically
describes the data collection and annotation of the corpus. An
annotation scheme has been proposed for the annotation of emotion-
related information including the emotion type, the emotion cause,
the emotion reaction, the use of rhetorical question, the opinion
target (i.e. the semantic role in an event that triggers an
emotion), etc. Corpus data shows that the annotated items are of
great value to the identification of implicit emotions. We believe
that the corpus will be a useful resource for both explicit and
implicit emotion classification and detection as well as event
classification.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.203}
}

@InProceedings{debruyne-declercq-hoste:2020:LREC,
  author    = {De Bruyne, Luna  and  De Clercq, Orphee  and  Hoste,
Veronique},
  title     = {An Emotional Mess! Deciding on a Framework for
Building a Dutch Emotion-Annotated Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1643--1651},
  abstract  = {Seeing the myriad of existing emotion models, with
the categorical versus dimensional opposition the most important
dividing line, building an emotion-annotated corpus requires some
well thought-out strategies concerning framework choice. In our work
on automatic emotion detection in Dutch texts, we investigate this
problem by means of two case studies. We find that the labels joy,
love, anger, sadness and fear are well-suited to annotate texts
coming from various domains and topics, but that the connotation of
the labels strongly depends on the origin of the texts. Moreover, it
seems that information is lost when an emotional state is forcedly
classified in a limited set of categories, indicating that a bi-
representational format is desirable when creating an emotion
corpus.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.204}
}

@InProceedings{haider-EtAl:2020:LREC,
  author    = {Haider, Thomas  and  Eger, Steffen  and  Kim, Evgeny
and  Klinger, Roman  and  Menninghaus, Winfried},
  title     = {PO-EMO: Conceptualization, Annotation, and Modeling
of Aesthetic Emotions in German and English Poetry},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1652--1663},
  abstract  = {Most approaches to emotion analysis of social media,
literature, news, and other domains focus exclusively on basic
emotion categories as defined by Ekman or Plutchik. However, art
(such as literature) enables engagement in a broader range of more
complex and subtle emotions. These have been shown to also include
mixed emotional responses. We consider emotions in poetry as they
are elicited in the reader, rather than what is expressed in the
text or intended by the author. Thus, we conceptualize a set of
aesthetic emotions that are predictive of aesthetic appreciation in
the reader, and allow the annotation of multiple labels per line to
capture mixed emotions within their context. We evaluate this novel
setting in an annotation experiment both with carefully trained
experts and via crowdsourcing. Our annotation with experts leads to
an acceptable agreement of k = .70, resulting in a consistent
dataset for future large scale analysis. Finally, we conduct first
emotion classification experiments based on BERT, showing that
identifying aesthetic emotions is challenging in our data, with up
to .52 F1-micro on the German subset. Data and resources are
available at https://github.com/tnhaider/poetry-emotion.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.205}
}

@InProceedings{sedoc-EtAl:2020:LREC,
```

```
  author    = {Sedoc, João  and  Buechel, Sven  and  Nachmany,
Yehonathan  and  Buffone, Anneke  and  Ungar, Lyle},
  title     = {Learning Word Ratings for Empathy and Distress from
Document-Level User Responses},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1664--1673},
  abstract  = {Despite the excellent performance of black box
approaches to modeling sentiment and emotion, lexica (sets of
informative words and associated weights) that characterize
different emotions are indispensable to the NLP community because
they allow for interpretable and robust predictions. Emotion
analysis of text is increasing in popularity in NLP; however,
manually creating lexica for psychological constructs such as
empathy has proven difficult. This paper automatically creates
empathy word ratings from document-level ratings. The underlying
problem of learning word ratings from higher-level supervision has
to date only been addressed in an ad hoc fashion and has not used
deep learning methods. We systematically compare a number of
approaches to learning word ratings from higher-level supervision
against a Mixed-Level Feed Forward Network (MLFFN), which we find
performs best, and use the MLFFN to create the first-ever empathy
lexicon. We then use Signed Spectral Clustering to gain insights
into the resulting words. The empathy and distress lexica are
publicly available at: http://www.wwbp.org/lexica.html.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.206}
}

@InProceedings{dadas-perekiewicz-powiata:2020:LREC,
  author    = {Dadas, Slawomir  and  Perełkiewicz, Michał  and
Poświata, Rafał},
  title     = {Evaluation of Sentence Representations in Polish},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1674--1680},
  abstract  = {Methods for learning sentence representations have
been actively developed in recent years. However, the lack of pre-
trained models and datasets annotated at the sentence level has been
a problem for low-resource languages such as Polish which led to
less interest in applying these methods to language-specific tasks.
In this study, we introduce two new Polish datasets for evaluating
sentence embeddings and provide a comprehensive evaluation of eight
sentence representation methods including Polish and multilingual
models. We consider classic word embedding models, recently
developed contextual embeddings and multilingual sentence encoders,
showing strengths and weaknesses of specific approaches. We also
```

examine different methods of aggregating word vectors into a single sentence vector.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.207}
}

@InProceedings{riad-EtAl:2020:LREC,
  author   = {Riad, Rachid and Bachoud-Lévi, Anne-Catherine and Rudzicz, Frank and Dupoux, Emmanuel},
  title   = {Identification of Primary and Collateral Tracks in Stuttered Speech},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1681--1688},
  abstract = {Disfluent speech has been previously addressed from two main perspectives: the clinical perspective focusing on diagnostic, and the Natural Language Processing (NLP) perspective aiming at modeling these events and detect them for downstream tasks. In addition, previous works often used different metrics depending on whether the input features are text or speech, making it difficult to compare the different contributions. Here, we introduce a new evaluation framework for disfluency detection inspired by the clinical and NLP perspective together with the theory of performance from (Clark, 1996) which distinguishes between primary and collateral tracks. We introduce a novel forced-aligned disfluency dataset from a corpus of semi-directed interviews, and present baseline results directly comparing the performance of text-based features (word and span information) and speech-based (acoustic-prosodic information). Finally, we introduce new audio features inspired by the word-based span features. We show experimentally that using these features outperformed the baselines for speech-based predictions on the present dataset.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.208}
}

@InProceedings{ghio-EtAl:2020:LREC,
  author   = {Ghio, Alain and Lalain, Muriel and Giusti, Laurence and Fredouille, Corinne and Woisard, Virginie},
  title   = {How to Compare Automatically Two Phonological Strings: Application to Intelligibility Measurement in the Case of Atypical Speech},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1689--1694},
  abstract = {Atypical speech productions, regardless of their origins (accents, learning, pathology), need to be assessed with regard to "typical" or "expected" productions. Evaluation is

necessarily based on comparisons between linguistic forms produced and linguistic forms expected. In the field of speech disorders, the intelligibility of a patient is evaluated in order to measure the functional impact of his/her pathology on his/her oral communication. The usual method is to transcribe orthographic linguistic forms perceived and to assign a global and imprecise rating based on their correctness or incorrect. To obtain a more precise evaluation of the production deviations, we propose a measurement method based on phonological transcriptions. An algorithm computes automatically and finely the distances between the phonological forms produced and expected from cost matrices based on the differences of features between phonemes. A first test of this method among a large population of healthy speakers and patients treated for cancer of the oral and pharyngeal cavities has proved its validity.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.209}
}

@InProceedings{schwab-EtAl:2020:LREC,
  author   = {Schwab, Didier and Trial, Pauline and Vaschalde, Céline and Vial, Loïc and Esperanca-Rodier, Emmanuelle and Lecouteux, Benjamin},
  title   = {Providing Semantic Knowledge to a Set of Pictograms for People with Disabilities: a Set of Links between WordNet and Arasaac: Arasaac-WN},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {166--171},
  abstract = {This article presents a resource that links WordNet, the widely known lexical and semantic database, and Arasaac, the largest freely available database of pictograms. Pictograms are a tool that is more and more used by people with cognitive or communication disabilities. However, they are mainly used manually via workbooks, whereas caregivers and families would like to use more automated tools (use speech to generate pictograms, for example). In order to make it possible to use pictograms automatically in NLP applications, we propose a database that links them to semantic knowledge. This resource is particularly interesting for the creation of applications that help people with cognitive disabilities, such as text-to-picto, speech-to-picto, picto-to-speech... In this article, we explain the needs for this database and the problems that have been identified. Currently, this resource combines approximately 800 pictograms with their corresponding WordNet synsets and it is accessible both through a digital collection and via an SQL database. Finally, we propose a method with associated tools to make our resource language-independent: this method was applied to create a first text-to-picto prototype for the French language. Our resource is distributed freely under a Creative Commons license at the following URL: https://github.com/getalp/Arasaac-WN.},

  url   = {https://www.aclweb.org/anthology/2020.lrec-1.21}
}

@InProceedings{liu-zeng-li:2020:LREC,
  author   = {Liu, Sennan and Zeng, Shuang and Li, Sujian},
  title   = {Evaluating Text Coherence at Sentence and Paragraph Levels},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1695--1703},
  abstract = {In this paper, to evaluate text coherence, we propose the paragraph ordering task as well as conducting sentence ordering. We collected four distinct corpora from different domains on which we investigate the adaptation of existing sentence ordering methods to a paragraph ordering task. We also compare the learnability and robustness of existing models by artificially creating mini datasets and noisy datasets respectively and verifying the efficiency of established models under these circumstances. Furthermore, we carry out human evaluation on the rearranged passages from two competitive models and confirm that WLCS-l is a better metric performing significantly higher correlations with human rating than τ , the most prevalent metric used before. Results from these evaluations show that except for certain extreme conditions, the recurrent graph neural network-based model is an optimal choice for coherence modeling.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.210}
}

@InProceedings{berniercolborne-langlais:2020:LREC,
  author   = {Bernier-Colborne, Gabriel and Langlais, Phillippe},
  title   = {HardEval: Focusing on Challenging Tokens to Assess Robustness of NER},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1704--1711},
  abstract = {To assess the robustness of NER systems, we propose an evaluation method that focuses on subsets of tokens that represent specific sources of errors: unknown words and label shift or ambiguity. These subsets provide a system-agnostic basis for evaluating specific sources of NER errors and assessing room for improvement in terms of robustness. We analyze these subsets of challenging tokens in two widely-used NER benchmarks, then exploit them to evaluate NER systems in both in-domain and out-of-domain settings. Results show that these challenging tokens explain the majority of errors made by modern NER systems, although they represent only a small fraction of test tokens. They also indicate

that label shift is harder to deal with than unknown words, and that there is much more room for improvement than the standard NER evaluation procedure would suggest. We hope this work will encourage NLP researchers to adopt rigorous and meaningful evaluation methods, and will help them develop more robust models.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.211}
}

@InProceedings{iwatsuki-boudin-aizawa:2020:LREC,
  author   = {Iwatsuki, Kenichi and Boudin, Florian and Aizawa, Akiko},
  title   = {An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1712--1720},
  abstract = {Formulaic expressions, such as ‘in this paper we propose’, are used by authors of scholarly papers to perform communicative functions; the communicative function of the present example is ‘stating the aim of the paper’. Collecting such expressions and pairing them with their communicative functions would be highly valuable for various tasks, particularly for writing assistance. However, such collection and paring in a principled and automated manner would require high-quality annotated data, which are not available. In this study, we address this shortcoming by creating a manually annotated dataset for detecting communicative functions in sentences. Starting from a seed list of labelled formulaic expressions, we retrieved new sentences from scholarly papers in the ACL Anthology and asked multiple human evaluators to label communicative functions. To show the usefulness of our dataset, we conducted a series of experiments that determined to what extent sentence representations acquired by recent models, such as word2vec and BERT, can be employed to detect communicative functions in sentences.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.212}
}

@InProceedings{fassetti-fassetti:2020:LREC,
  author   = {Fassetti, Fabio and Fassetti, Ilaria},
  title   = {An Automatic Tool For Language Evaluation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1721--1726},
  abstract = {The aim of evaluating children speech and language is to measure their communication skills. In particular, the speech language pathologist is interested in determining the child's

```
impairments in the areas of language, articulation, voice, fluency
and swallowing. In literature some standardized tests have been
proposed to assess and screen developmental language impairments but
they require manual laborious transcription, annotation and
calculation. This work is very time demanding and, also, may
introduce several kinds of errors in the evaluation phase and non-
uniform evaluations. In order to help therapists, a system
performing automated evaluation is proposed. Providing as input the
correct sentence and the sentence produced by patients, the
technique evaluates the level of the verbal production and returns a
score. The main phases of the method concern an ad-hoc
transformation of the produced sentence in the reference sentence
and in the evaluation of the cost of this transformation. Since the
cost function is related to many weights, a learning phase is
defined to automatically set such weights.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.213}
}

@InProceedings{boydgraber-EtAl:2020:LREC,
  author    = {Boyd-Graber, Jordan  and  Guo, Fenfei  and
Findlater, Leah  and  Iyyer, Mohit},
  title     = {Which Evaluations Uncover Sense Representations that
Actually Make Sense?},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1727--1738},
  abstract  = {Text representations are critical for modern natural
language processing. One form of text representation, sense-specific
embeddings, reflect a word's sense in a sentence better than single-
prototype word embeddings tied to each type. However, existing sense
representations are not uniformly better: although they work well
for computer-centric evaluations, they fail for human-centric tasks
like inspecting a language's sense inventory. To expose this
discrepancy, we propose a new coherence evaluation for sense
embeddings. We also describe a minimal model (Gumbel Attention for
Sense Induction) optimized for discovering interpretable sense
representations that are more coherent than existing sense
embeddings.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.214}
}

@InProceedings{lai-EtAl:2020:LREC,
  author    = {Lai, Yi-An  and  Zhu, Xuan  and  Zhang, Yi  and
Diab, Mona},
  title     = {Diversity, Density, and Homogeneity: Quantitative
Characteristic Metrics for Text Collections},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1739--1746},
  abstract  = {Summarizing data samples by quantitative measures has
a long history, with descriptive statistics being a case in point.
However, as natural language processing methods flourish, there are
still insufficient characteristic metrics to describe a collection
of texts in terms of the words, sentences, or paragraphs they
comprise. In this work, we propose metrics of diversity, density,
and homogeneity that quantitatively measure the dispersion,
sparsity, and uniformity of a text collection. We conduct a series
of simulations to verify that each metric holds desired properties
and resonates with human intuitions. Experiments on real-world
datasets demonstrate that the proposed characteristic metrics are
highly correlated with text classification performance of a renowned
model, BERT, which could inspire future applications.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.215}
}

@InProceedings{min-chan-zhao:2020:LREC,
  author    = {Min, Bonan  and  Chan, Yee Seng  and  Zhao, Lingjun},
  title     = {Towards Few-Shot Event Mention Retrieval: An
Evaluation Framework and A Siamese Network Approach},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1747--1752},
  abstract  = {Automatically analyzing events in a large amount of
text is crucial for situation awareness and decision making.
Previous approaches treat event extraction as "one size fits all"
with an ontology defined a priori. The resulted extraction models
are built just for extracting those types in the ontology. These
approaches cannot be easily adapted to new event types nor new
domains of interest. To accommodate personalized event-centric
information needs, this paper introduces the few-shot Event Mention
Retrieval (EMR) task: given a user-supplied query consisting of a
handful of event mentions, return relevant event mentions found in a
corpus. This formulation enables "query by example", which
drastically lowers the bar of specifying event-centric information
needs. The retrieval setting also enables fuzzy search. We present
an evaluation framework leveraging existing event datasets such as
ACE. We also develop a Siamese Network approach, and show that it
performs better than ad-hoc retrieval models in the few-shot EMR
setting.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.216}
}

@InProceedings{horbach-EtAl:2020:LREC,
  author    = {Horbach, Andrea  and  Aldabe, Itziar  and  Bexte,
Marie  and  Lopez de Lacalle, Oier  and  Maritxalar, Montse},
  title     = {Linguistic Appropriateness and Pedagogic Usefulness
```

of Reading Comprehension Questions},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1753--1762},
  abstract = {Automatic generation of reading comprehension questions is a topic receiving growing interest in the NLP community, but there is currently no consensus on evaluation metrics and many approaches focus on linguistic quality only while ignoring the pedagogic value and appropriateness of questions. This paper overcomes such weaknesses by a new evaluation scheme where questions from the questionnaire are structured in a hierarchical way to avoid confronting human annotators with evaluation measures that do not make sense for a certain question. We show through an annotation study that our scheme can be applied, but that expert annotators with some level of expertise are needed. We also created and evaluated two new evaluation data sets from the biology domain for Basque and German, composed of questions written by people with an educational background, which will be publicly released. Results show that manually generated questions are in general both of higher linguistic as well as pedagogic quality and that among the human generated questions, teacher-generated ones tend to be most useful.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.217}
}

@InProceedings{born-bacher-markert:2020:LREC,
  author   = {Born, Leo and Bacher, Maximilian and Markert, Katja},
  title   = {Dataset Reproducibility and IR Methods in Timeline Summarization},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1763--1771},
  abstract = {Timeline summarization (TLS) generates a dated overview of real-world events based on event-specific corpora. The two standard datasets for this task were collected using Google searches for news reports on given events. Not only is this IR method not reproducible at different search times, it also uses components (such as document popularity) that are not always available for any large news corpus. It is unclear how TLS algorithms fare when provided with event corpora collected with varying IR methods. We therefore construct event-specific corpora from a large static background corpus, the newsroom dataset, using differing, relatively simple IR methods based on raw text alone. We show that the choice of IR method plays a crucial role in the performance of various TLS algorithms. A weak TLS algorithm can even

match a stronger one by employing a stronger IR method in the data collection phase. Furthermore, the results of TLS systems are often highly sensitive to additional sentence filtering. We consequently advocate for integrating IR into the development of TLS systems and having a common static background corpus for evaluation of TLS systems.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.218}
}

@InProceedings{nadig-braschler-stockinger:2020:LREC,
  author   = {Nadig, Stefanie and Braschler, Martin and Stockinger, Kurt},
  title   = {Database Search vs. Information Retrieval: A Novel Method for Studying Natural Language Querying of Semi-Structured Data},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1772--1779},
  abstract = {The traditional approach of querying a relational database is via a formal language, namely SQL. Recent developments in the design of natural language interfaces to databases show promising results for querying either with keywords or with full natural language queries and thus render relational databases more accessible to non-tech savvy users. Such enhanced relational databases basically use a search paradigm which is commonly used in the field of information retrieval. However, the way systems are evaluated in the database and the information retrieval communities often differs due to a lack of common benchmarks. In this paper, we provide an adapted benchmark data set that is based on a test collection originally used to evaluate information retrieval systems. The data set contains 45 information needs developed on the Internet Movie Database (IMDb), including corresponding relevance assessments. By mapping this benchmark data set to a relational database schema, we enable a novel way of directly comparing database search techniques with information retrieval. To demonstrate the feasibility of our approach, we present an experimental evaluation that compares SODA, a keyword-enabled relational database system, against the Terrier information retrieval system and thus lays the foundation for a future discussion of evaluating database systems that support natural language interfaces.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.219}
}

@InProceedings{tulkens-sandra-daelemans:2020:LREC,
  author   = {Tulkens, Stéphan and Sandra, Dominiek and Daelemans, Walter},
  title   = {Orthographic Codes and the Neighborhood Effect: Lessons from Information Theory},
  booktitle   = {Proceedings of The 12th Language Resources and

Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {172--181},
  abstract = {We consider the orthographic neighborhood effect: the effect that words with more orthographic similarity to other words are read faster. The neighborhood effect serves as an important control variable in psycholinguistic studies of word reading, and explains variance in addition to word length and word frequency. Following previous work, we model the neighborhood effect as the average distance to neighbors in feature space for three feature sets: slots, character ngrams and skipgrams. We optimize each of these feature sets and find evidence for language-independent optima, across five megastudy corpora from five alphabetic languages. Additionally, we show that weighting features using the inverse of mutual information (MI) improves the neighborhood effect significantly for all languages. We analyze the inverse feature weighting, and show that, across languages, grammatical morphemes get the lowest weights. Finally, we perform the same experiments on Korean Hangul, a non-alphabetic writing system, where we find the opposite results: slower responses as a function of denser neighborhoods, and a negative effect of inverse feature weighting. This raises the question of whether this is a cognitive effect, or an effect of the way we represent Hangul orthography, and indicates more research is needed.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.22}
}

@InProceedings{grimsley-mayfield-rsbursten:2020:LREC,
  author   = {Grimsley, Christopher and Mayfield, Elijah and R.S. Bursten, Julia},
  title   = {Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1780--1790},
  abstract = {As the demand for explainable deep learning grows in the evaluation of language technologies, the value of a principled grounding for those explanations grows as well. Here we study the state-of-the-art in explanation for neural models for NLP tasks from the viewpoint of philosophy of science. We focus on recent evaluation work that finds brittleness in explanations obtained through attention mechanisms. We harness philosophical accounts of explanation to suggest broader conclusions from these studies. From this analysis, we assert the impossibility of causal explanations from attention layers over text data. We then introduce NLP researchers to contemporary philosophy of science theories that allow robust yet non-causal reasoning in explanation, giving

computer scientists a vocabulary for future research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.220}
}

@InProceedings{marczyk-EtAl:2020:LREC,
  author   = {Marczyk, Anna and Ghio, Alain and Lalain, Muriel and Rebourg, Marie and Fredouille, Corinne and Woisard, Virginie},
  title   = {Have a Cake and Eat it Too: Assessing Discriminating Performance of an Intelligibility Index Obtained from a Reduced Sample Size},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1791--1795},
  abstract = {This paper investigates random vs. phonetically motivated reduction of linguistic material used in an intelligibility task in speech disordered populations and the subsequent impact on the discrimination classifier quantified by the area under the receiver operating characteristics curve (AUC of ROC). The comparison of obtained accuracy indexes shows that when the sample size is reduced based on a phonetic criterium—here, related to phonotactic complexity—, the classifier has a higher ranking ability than when the linguistic material is arbitrarily reduced. Crucially, downsizing the linguistic sample to about 30\% of the original dataset does not diminish the discriminatory performance of the classifier. This result is of significant interest to both clinicians and patients as it validates a tool that is both reliable and efficient.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.221}
}

@InProceedings{moeed-EtAl:2020:LREC2,
  author   = {Moeed, Abdul and An, Yang and Hagerer, Gerhard and Groh, Georg},
  title   = {Evaluation Metrics for Headline Generation Using Deep Pre-Trained Embeddings},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1796--1802},
  abstract = {With the explosive growth in textual data, it is becoming increasingly important to summarize text automatically. Recently, generative language models have shown promise in abstractive text summarization tasks. Since these models rephrase text and thus use similar but different words as found in the summarized text, existing metrics such as ROUGE that use n-gram overlap may not be optimal. Therefore we evaluate two embedding-

based evaluation metrics that are applicable to abstractive summarization: Fr´echet embedding distance, which has been introduced recently, and angular embedding similarity, which is our proposed metric. To demonstrate the utility of both metrics, we analyze the headline generation capacity of two state-of-the-art language models: GPT-2 and ULMFiT. In particular, our proposed metric shows close relation with human judgments in our experiments and has overall better correlations with them. To provide reproducibility, the source code plus human assessments of our experiments is available on GitHub.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.222}
}

@InProceedings{aguilar-kar-solorio:2020:LREC,
  author   = {Aguilar, Gustavo and Kar, Sudipta and Solorio, Thamar},
  title   = {LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1803--1813},
  abstract = {Recent trends in NLP research have raised an interest in linguistic code-switching (CS); modern approaches have been proposed to solve a wide range of NLP tasks on multiple language pairs. Unfortunately, these proposed methods are hardly generalizable to different code-switched languages. In addition, it is unclear whether a model architecture is applicable for a different task while still being compatible with the code-switching setting. This is mainly because of the lack of a centralized benchmark and the sparse corpora that researchers employ based on their specific needs and interests. To facilitate research in this direction, we propose a centralized benchmark for Linguistic Code-switching Evaluation (LinCE) that combines eleven corpora covering four different code-switched language pairs (i.e., Spanish-English, Nepali-English, Hindi-English, and Modern Standard Arabic-Egyptian Arabic) and four tasks (i.e., language identification, named entity recognition, part-of-speech tagging, and sentiment analysis). As part of the benchmark centralization effort, we provide an online platform where researchers can submit their results while comparing with others in real-time. In addition, we provide the scores of different popular models, including LSTM, ELMo, and multilingual BERT so that the NLP community can compare against state-of-the-art systems. LinCE is a continuous effort, and we will expand it with more low-resource languages and tasks.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.223}
}

@InProceedings{sjblom-creutz-scherrer:2020:LREC,
  author   = {Sjöblom, Eetu and Creutz, Mathias and Scherrer, Yves},

```
  title     = {Paraphrase Generation and Evaluation on Colloquial-
Style Sentences},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1814--1822},
  abstract  = {In this paper, we investigate paraphrase generation
in the colloquial domain. We use state-of-the-art neural machine
translation models trained on the Opusparcus corpus to generate
paraphrases in six languages: German, English, Finnish, French,
Russian, and Swedish. We perform experiments to understand how data
selection and filtering for diverse paraphrase pairs affects the
generated paraphrases. We compare two different model architectures,
an RNN and a Transformer model, and find that the Transformer does
not generally outperform the RNN. We also conduct human evaluation
on five of the six languages and compare the results to the
automatic evaluation metrics BLEU and the recently proposed
BERTScore. The results advance our understanding of the trade-offs
between the quality and novelty of generated paraphrases, affected
by the data selection method. In addition, our comparison of the
evaluation methods shows that while BLEU correlates well with human
judgments at the corpus level, BERTScore outperforms BLEU in both
corpus and sentence-level evaluation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.224}
}

@InProceedings{han-hayashi-miyao:2020:LREC,
  author    = {Han, Namgi  and  Hayashi, Katsuhiko  and  Miyao,
Yusuke},
  title     = {Analyzing Word Embedding Through Structural Equation
Modeling},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1823--1832},
  abstract  = {Many researchers have tried to predict the accuracies
of extrinsic evaluation by using intrinsic evaluation to evaluate
word embedding. The relationship between intrinsic and extrinsic
evaluation, however, has only been studied with simple correlation
analysis, which has difficulty capturing complex cause-effect
relationships and integrating external factors such as the
hyperparameters of word embedding. To tackle this problem, we employ
partial least squares path modeling (PLS-PM), a method of structural
equation modeling developed for causal analysis. We propose a causal
diagram consisting of the evaluation results on the BATS, VecEval,
and SentEval datasets, with a causal hypothesis that linguistic
knowledge encoded in word embedding contributes to solving
downstream tasks. Our PLS-PM models are estimated with 600 word
```

```
embeddings, and we prove the existence of causal relations between
linguistic knowledge evaluated on BATS and the accuracies of
downstream tasks evaluated on VecEval and SentEval in our PLS-PM
models. Moreover, we show that the PLS-PM models are useful for
analyzing the effect of hyperparameters, including the training
algorithm, corpus, dimension, and context window, and for validating
the effectiveness of intrinsic evaluation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.225}
}

@InProceedings{prokopalo-EtAl:2020:LREC,
  author    = {Prokopalo, Yevhenii  and  Meignier, Sylvain  and
Galibert, Olivier  and  Barrault, Loic  and  Larcher, Anthony},
  title     = {Evaluation of Lifelong Learning Systems},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1833--1841},
  abstract  = {Current intelligent systems need the expensive
support of machine learning experts to sustain their performance
level when used on a daily basis. To reduce this cost, i.e.
remaining free from any machine learning expert, it is reasonable to
implement lifelong (or continuous) learning intelligent systems that
will continuously adapt their model when facing changing execution
conditions. In this work, the systems are allowed to refer to human
domain experts who can provide the system with relevant knowledge
about the task. Nowadays, the fast growth of lifelong learning
systems development rises the question of their evaluation. In this
article we propose a generic evaluation methodology for the specific
case of lifelong learning systems. Two steps will be considered.
First, the evaluation of human-assisted learning (including active
and/or interactive learning) outside the context of lifelong
learning. Second, the system evaluation across time, with
propositions of how a lifelong learning intelligent system should be
evaluated when including human assisted learning or not.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.226}
}

@InProceedings{hajnicz:2020:LREC,
  author    = {Hajnicz, Elżbieta},
  title     = {Interannotator Agreement for Lexico-Semantic
Annotation of a Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1842--1848},
  abstract  = {This paper examines the procedure for lexico-semantic
annotation of the Basic Corpus of Polish Metaphors that is the first
```

step for annotating metaphoric expressions occurring in it. The procedure involves correcting the morphosyntactic annotation of part of the corpus that is automatically annotated on the morphosyntactic level. The main procedure concerns annotation of adjectives, adverbs, nouns and verbs (including gerunds and participles), including abbreviations of the words that belong to the above classes. It is composed of three steps: deciding whether a particular occurrence of a word is asemantic (e.g. anaphoric or strictly grammatical), whether we are dealing with a multi-word expression, reciprocal usages of the się marker and pluralia tantum, which may involve annotation with two lexical units (having two different lemmas) for a single token. We propose an interannotator agreement statistics adequate for this procedure. Finally, we discuss the preliminary results of annotation of a fragment of the corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.227}
}

@InProceedings{nther:2020:LREC,
  author   = {Näther, Markus},
  title   = {An In-Depth Comparison of 14 Spelling Correction Tools on a Common Benchmark},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1849--1857},
  abstract = {Determining and correcting spelling and grammar errors in text is an important but surprisingly difficult task. There are several reasons why this remains challenging. Errors may consist of simple typing errors like deleted, substituted, or wrongly inserted letters, but may also consist of word confusions where a word was replaced by another one. In addition, words may be erroneously split into two parts or get concatenated. Some words can contain hyphens, because they were split at the end of a line or are compound words with a mandatory hyphen. In this paper, we provide an extensive evaluation of 14 spelling correction tools on a common benchmark. In particular, the evaluation provides a detailed comparison with respect to 12 error categories. The benchmark consists of sentences from the English Wikipedia, which were distorted using a realistic error model. Measuring the quality of an algorithm with respect to these error categories requires an alignment of the original text, the distorted text and the corrected text provided by the tool. We make our benchmark generation and evaluation tools publicly available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.228}
}

@InProceedings{yuan-sharoff:2020:LREC,
  author   = {Yuan, Yu and Sharoff, Serge},
  title   = {Sentence Level Human Translation Quality Estimation with Attention-based Neural Networks},

  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1858--1865},
  abstract = {This paper explores the use of Deep Learning methods for automatic estimation of quality of human translations. Automatic estimation can provide useful feedback for translation teaching, examination and quality control. Conventional methods for solving this task rely on manually engineered features and external knowledge. This paper presents an end-to-end neural model without feature engineering, incorporating a cross attention mechanism to detect which parts in sentence pairs are most relevant for assessing quality. Another contribution concerns oprediction of fine-grained scores for measuring different aspects of translation quality, such as terminological accuracy or idiomatic writing. Empirical results on a large human annotated dataset show that the neural model outperforms feature-based methods significantly. The dataset and the tools are available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.229}
}

@InProceedings{kerz-EtAl:2020:LREC,
  author   = {Kerz, Elma and Pruneri, Fabio and Wiechmann, Daniel and Qiao, Yu and Ströbel, Marcus},
  title   = {Understanding the Dynamics of Second Language Writing through Keystroke Logging and Complexity Contours},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {182--188},
  abstract = {The purpose of this paper is twofold: [1] to introduce, to our knowledge, the largest available resource of keystroke logging (KSL) data generated by Etherpad (https://etherpad.org/), an open-source, web-based collaborative real-time editor, that captures the dynamics of second language (L2) production and [2] to relate the behavioral data from KSL to indices of syntactic and lexical complexity of the texts produced obtained from a tool that implements a sliding window approach capturing the progression of complexity within a text. We present the procedures and measures developed to analyze a sample of 14,913,009 keystrokes in 3,454 texts produced by 512 university students (upper-intermediate to advanced L2 learners of English) (95,354 sentences and 18,32,027 words) aiming to achieve a better alignment between keystroke-logging measures and underlying cognitive processes, on the one hand, and L2 writing performance measures, on the other hand. The resource introduced in this paper is a reflection of increasing recognition of the urgent need to obtain ecologically valid data that have the potential to transform our current

understanding of mechanisms underlying the development of literacy (reading and writing) skills.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.23}
}

@InProceedings{alves-thakkar-tadi:2020:LREC,
  author   = {Alves, Diego and Thakkar, Gaurish and Tadić, Marko},
  title   = {Evaluating Language Tools for Fifteen EU-official Under-resourced Languages},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1866--1873},
  abstract = {This article presents the results of the evaluation campaign of language tools available for fifteen EU-official under-resourced languages. The evaluation was conducted within the MSC ITN CLEOPATRA action that aims at building the cross-lingual event-centric knowledge processing on top of the application of linguistic processing chains (LPCs) for at least 24 EU-official languages. In this campaign, we concentrated on three existing NLP platforms (Stanford CoreNLP, NLP Cube, UDPipe) that all provide models for under-resourced languages and in this first run we covered 15 under-resourced languages for which the models were available. We present the design of the evaluation campaign and present the results as well as discuss them. We considered the difference between reported and our tested results within a single percentage point as being within the limits of acceptable tolerance and thus consider this result as reproducible. However, for a number of languages, the results are below what was reported in the literature, and in some cases, our testing results are even better than the ones reported previously. Particularly problematic was the evaluation of NERC systems. One of the reasons is the absence of universally or cross-lingually applicable named entities classification scheme that would serve the NERC task in different languages analogous to the Universal Dependency scheme in parsing task. To build such a scheme has become one of our the future research directions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.230}
}

@InProceedings{lakmal-EtAl:2020:LREC,
  author   = {Lakmal, Dimuthu and Ranathunga, Surangika and Peramuna, Saman and Herath, Indu},
  title   = {Word Embedding Evaluation for Sinhala},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1874--1881},

  abstract = {This paper presents the first ever comprehensive evaluation of different types of word embeddings for Sinhala language. Three standard word embedding models, namely, Word2Vec (both Skipgram and CBOW), FastText, and Glove are evaluated under two types of evaluation methods: intrinsic evaluation and extrinsic evaluation. Word analogy and word relatedness evaluations were performed in terms of intrinsic evaluation, while sentiment analysis and part-of-speech (POS) tagging were conducted as the extrinsic evaluation tasks. Benchmark datasets used for intrinsic evaluations were carefully crafted considering specific linguistic features of Sinhala. In general, FastText word embeddings with 300 dimensions reported the finest accuracies across all the evaluation tasks, while Glove reported the lowest results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.231}
}

@InProceedings{aspillaga-carvallo-araujo:2020:LREC,
  author   = {Aspillaga, Carlos and Carvallo, Andrés and Araujo, Vladimir},
  title   = {Stress Test Evaluation of Transformer-based Models in Natural Language Understanding Tasks},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1882--1894},
  abstract = {There has been significant progress in recent years in the field of Natural Language Processing thanks to the introduction of the Transformer architecture. Current state-of-the-art models, via a large number of parameters and pre-training on massive text corpus, have shown impressive results on several downstream tasks. Many researchers have studied previous (non-Transformer) models to understand their actual behavior under different scenarios, showing that these models are taking advantage of clues or failures of datasets and that slight perturbations on the input data can severely reduce their performance. In contrast, recent models have not been systematically tested with adversarial-examples in order to show their robustness under severe stress conditions. For that reason, this work evaluates three Transformer-based models (RoBERTa, XLNet, and BERT) in Natural Language Inference (NLI) and Question Answering (QA) tasks to know if they are more robust or if they have the same flaws as their predecessors. As a result, our experiments reveal that RoBERTa, XLNet and BERT are more robust than recurrent neural network models to stress tests for both NLI and QA tasks. Nevertheless, they are still very fragile and demonstrate various unexpected behaviors, thus revealing that there is still room for future improvement in this field.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.232}
}

@InProceedings{janz-EtAl:2020:LREC,

```
  author    = {Janz, Arkadiusz  and  Kopociński, Łukasz  and
Piasecki, Maciej  and  Pluwak, Agnieszka},
  title     = {Brand-Product Relation Extraction Using Heterogeneous
Vector Space Representations},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1895--1901},
  abstract  = {Relation Extraction is a fundamental NLP task. In
this paper we investigate the impact of underlying text
representation on the performance of neural classification models in
the task of Brand-Product relation extraction. We also present the
methodology of preparing annotated textual corpora for this task and
we provide valuable insight into the properties of Brand-Product
relations existing in textual corpora. The problem is approached
from a practical angle of applications Relation Extraction in
facilitating commercial Internet monitoring.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.233}
}

@InProceedings{buljan-EtAl:2020:LREC,
  author    = {Buljan, Maja  and  Nivre, Joakim  and  Oepen, Stephan
and  Øvrelid, Lilja},
  title     = {A Tale of Three Parsers: Towards Diagnostic
Evaluation for Meaning Representation Parsing},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1902--1909},
  abstract  = {We discuss methodological choices in contrastive and
diagnostic evaluation in meaning representation parsing, i.e.
mapping from natural language utterances to graph-based encodings of
its semantic structure. Drawing inspiration from earlier work in
syntactic dependency parsing, we transfer and refine several
quantitative diagnosis techniques for use in the context of the 2019
shared task on Meaning Representation Parsing (MRP). As in parsing
proper, moving evaluation from simple rooted trees to general graphs
brings along its own range of challenges. Specifically, we seek to
begin to shed light on relative strenghts and weaknesses in
different broad families of parsing techniques. In addition to these
theoretical reflections, we conduct a pilot experiment on a
selection of top-performing MRP systems and one of the five meaning
representation frameworks in the shared task. Empirical results
suggest that the proposed methodology can be meaningfully applied to
parsing into graph-structured target representations, uncovering
hitherto unknown properties of the different systems that can inform
future development and cross-fertilization across approaches.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.234}
```

}

@InProceedings{yang-EtAl:2020:LREC,
  author   = {Yang, Mu and Chen, Chi-Yen and Lee, Yi-Hui and Zeng, Qian-hui and Ma, Wei-Yun and Shih, Chen-Yang and Chen, Wei-Jhih},
  title   = {Headword-Oriented Entity Linking: A Special Entity Linking Task with Dataset and Baseline},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1910--1917},
  abstract = {In this paper, we design headword-oriented entity linking (HEL), a specialized entity linking problem in which only the headwords of the entities are to be linked to knowledge bases; mention scopes of the entities do not need to be identified in the problem setting. This special task is motivated by the fact that in many articles referring to specific products, the complete full product names are rarely written; instead, they are often abbreviated to shorter, irregular versions or even just to their headwords, which are usually their product types, such as “stick” or “mask” in a cosmetic context. To fully design the special task, we construct a labeled cosmetic corpus as a public benchmark for this problem, and propose a product embedding model to address the task, where each product corresponds to a dense representation to encode the different information on products and their context jointly. Besides, to increase training data, we propose a special transfer learning framework in which distant supervision with heuristic patterns is first utilized, followed by supervised learning using a small amount of manually labeled data. The experimental results show that our model provides a strong benchmark performance on the special task.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.235}
}

@InProceedings{li-EtAl:2020:LREC,
  author   = {Li, Minghao and Cui, Lei and Huang, Shaohan and Wei, Furu and Zhou, Ming and Li, Zhoujun},
  title   = {TableBank: Table Benchmark for Image-based Table Detection and Recognition},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1918--1925},
  abstract = {We present TableBank, a new image-based table detection and recognition dataset built with novel weak supervision from Word and Latex documents on the internet. Existing research for image-based table detection and recognition usually fine-tunes pre-

trained models on out-of-domain data with a few thousand human-labeled examples, which is difficult to generalize on real-world applications. With TableBank that contains 417K high quality labeled tables, we build several strong baselines using state-of-the-art models with deep neural networks. We make TableBank publicly available and hope it will empower more deep learning approaches in the table detection and recognition task. The dataset and models can be downloaded from https://github.com/doc-analysis/TableBank.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.236}
}

@InProceedings{frej-schwab-chevallet:2020:LREC,
  author   = {Frej, Jibril and Schwab, Didier and Chevallet, Jean-Pierre},
  title   = {WIKIR: A Python Toolkit for Building a Large-scale Wikipedia-based English Information Retrieval Dataset},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1926--1933},
  abstract = {Over the past years, deep learning methods allowed for new state-of-the-art results in ad-hoc information retrieval. However such methods usually require large amounts of annotated data to be effective. Since most standard ad-hoc information retrieval datasets publicly available for academic research (e.g. Robust04, ClueWeb09) have at most 250 annotated queries, the recent deep learning models for information retrieval perform poorly on these datasets. These models (e.g. DUET, Conv-KNRM) are trained and evaluated on data collected from commercial search engines not publicly available for academic research which is a problem for reproducibility and the advancement of research. In this paper, we propose WIKIR: an open-source toolkit to automatically build large-scale English information retrieval datasets based on Wikipedia. WIKIR is publicly available on GitHub. We also provide wikIR59k: a large-scale publicly available dataset that contains 59,252 queries and 2,617,003 (query, relevant documents) pairs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.237}
}

@InProceedings{tanaka-EtAl:2020:LREC,
  author   = {Tanaka, Koji and Chu, Chenhui and Ren, Haolin and Renoust, Benjamin and Nakashima, Yuta and Takemura, Noriko and Nagahara, Hajime and Fujikawa, Takao},
  title   = {Constructing a Public Meeting Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1934--1940},

```
  abstract  = {In this paper, we propose a full pipeline of analysis
of a large corpus about a century of public meeting in historical
Australian news papers, from construction to visual exploration. The
corpus construction method is based on image processing and OCR. We
digitize and transcribe texts of the specific topic of public
meeting. Experiments show that our proposed method achieves a F-
score of 87.8\% for corpus construction. As a result, we built a
content search tool for temporal and semantic content analysis.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.238}
}

@InProceedings{kuniyoshi-EtAl:2020:LREC,
  author    = {Kuniyoshi, Fusataka  and  Makino, Kohei  and  Ozawa,
Jun  and  Miwa, Makoto},
  title     = {Annotating and Extracting Synthesis Process of All-
Solid-State Batteries from Scientific Literature},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1941--1950},
  abstract  = {The synthesis process is essential for achieving
computational experiment design in the field of inorganic materials
chemistry. In this work, we present a novel corpus of the synthesis
process for all-solid-state batteries and an automated machine
reading system for extracting the synthesis processes buried in the
scientific literature. We define the representation of the synthesis
processes using flow graphs, and create a corpus from the
experimental sections of 243 papers. The automated machine-reading
system is developed by a deep learning-based sequence tagger and
simple heuristic rule-based relation extractor. Our experimental
results demonstrate that the sequence tagger with the optimal
setting can detect the entities with a macro-averaged F1 score of
0.826, while the rule-based relation extractor can achieve high
performance with a macro-averaged F1 score of 0.887.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.239}
}

@InProceedings{oseki-asahara:2020:LREC,
  author    = {Oseki, Yohei  and  Asahara, Masayuki},
  title     = {Design of BCCWJ-EEG: Balanced Corpus with Human
Electroencephalography},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {189--194},
  abstract  = {The past decade has witnessed the happy marriage
between natural language processing (NLP) and the cognitive science
of language. Moreover, given the historical relationship between
```

biological and artificial neural networks, the advent of deep learning has re-sparked strong interests in the fusion of NLP and the neuroscience of language. Importantly, this inter-fertilization between NLP, on one hand, and the cognitive (neuro)science of language, on the other, has been driven by the language resources annotated with human language processing data. However, there remain several limitations with those language resources on annotations, genres, languages, etc. In this paper, we describe the design of a novel language resource called BCCWJ-EEG, the Balanced Corpus of Contemporary Written Japanese (BCCWJ) experimentally annotated with human electroencephalography (EEG). Specifically, after extensively reviewing the language resources currently available in the literature with special focus on eye-tracking and EEG, we summarize the details concerning (i) participants, (ii) stimuli, (iii) procedure, (iv) data preprocessing, (v) corpus evaluation, (vi) resource release, and (vii) compilation schedule. In addition, potential applications of BCCWJ-EEG to neuroscience and NLP will also be discussed.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.24}
}

@InProceedings{strobl-trabelsi-zaiane:2020:LREC,
  author   = {Strobl, Michael and Trabelsi, Amine and Zaiane, Osmar},
  title   = {WEXEA: Wikipedia EXhaustive Entity Annotation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1951--1958},
  abstract = {Building predictive models for information extraction from text, such as named entity recognition or the extraction of semantic relationships between named entities in text, requires a large corpus of annotated text. Wikipedia is often used as a corpus for these tasks where the annotation is a named entity linked by a hyperlink to its article. However, editors on Wikipedia are only expected to link these mentions in order to help the reader to understand the content, but are discouraged from adding links that do not add any benefit for understanding an article. Therefore, many mentions of popular entities (such as countries or popular events in history), or previously linked articles, as well as the article's entity itself, are not linked. In this paper, we discuss WEXEA, a Wikipedia EXhaustive Entity Annotation system, to create a text corpus based on Wikipedia with exhaustive annotations of entity mentions, i.e. linking all mentions of entities to their corresponding articles. This results in a huge potential for additional annotations that can be used for downstream NLP tasks, such as Relation Extraction. We show that our annotations are useful for creating distantly supervised datasets for this task. Furthermore, we publish all code necessary to derive a corpus from a raw Wikipedia dump, so that it can be reproduced by everyone.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.240}

}

@InProceedings{ferr-EtAl:2020:LREC,
  author   = {Ferré, Arnaud and Bossy, Robert and Ba, Mouhamadou and Deléger, Louise and Lavergne, Thomas and Zweigenbaum, Pierre and Nédellec, Claire},
  title   = {Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1959--1966},
  abstract = {Entity normalization (or entity linking) is an important subtask of information extraction that links entity mentions in text to categories or concepts in a reference vocabulary. Machine learning based normalization methods have good adaptability as long as they have enough training data per reference with a sufficient quality. Distributional representations are commonly used because of their capacity to handle different expressions with similar meanings. However, in specific technical and scientific domains, the small amount of training data and the relatively small size of specialized corpora remain major challenges. Recently, the machine learning-based CONTES method has addressed these challenges for reference vocabularies that are ontologies, as is often the case in life sciences and biomedical domains. And yet, its performance is dependent on manually annotated corpus. Furthermore, like other machine learning based methods, parametrization remains tricky. We propose a new approach to address the scarcity of training data that extends the CONTES method by corpus selection, pre-processing and weak supervision strategies, which can yield high-performance results without any manually annotated examples. We also study which hyperparameters are most influential, with sometimes different patterns compared to previous work. The results show that our approach significantly improves accuracy and outperforms previous state-of-the-art algorithms.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.241}
}

@InProceedings{bonin-EtAl:2020:LREC,
  author   = {Bonin, Francesca and Gleize, Martin and Finnerty, Ailbhe and Moore, Candice and Jochim, Charles and Norris, Emma and Hou, Yufang and Wright, Alison J. and Ganguly, Debasis and Hayes, Emily and Zink, Silje and Pascale, Alessandra and Mac Aonghusa, Pol and Michie, Susan},
  title   = {HBCP Corpus: A New Resource for the Analysis of Behavioural Change Intervention Reports},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1967--1975},
  abstract  = {Due to the fast pace at which research reports in
behaviour change are published, researchers, consultants and
policymakers would benefit from more automatic ways to process these
reports. Automatic extraction of the reports' intervention content,
population, settings and their results etc. are essential in
synthesising and summarising the literature. However, to the best of
our knowledge, no unique resource exists at the moment to facilitate
this synthesis. In this paper, we describe the construction of a
corpus of published behaviour change intervention evaluation reports
aimed at smoking cessation. We also describe and release the
annotation of 57 entities, that can be used as an off-the-shelf data
resource for tasks such as entity recognition, etc. Both the corpus
and the annotation dataset are being made available to the
community.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.242}
}

@InProceedings{lu-EtAl:2020:LREC,
  author    = {Lu, Di  and  Subburathinam, Ananya  and  Ji, Heng
and  May, Jonathan  and  Chang, Shih-Fu  and  Sil, Avi  and  Voss,
Clare},
  title     = {Cross-lingual Structure Transfer for Zero-resource
Event Extraction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {1976--1981},
  abstract  = {Most of the current cross-lingual transfer learning
methods for Information Extraction (IE) have been only applied to
name tagging. To tackle more complex tasks such as event extraction
we need to transfer graph structures (event trigger linked to
multiple arguments with various roles) across languages. We develop
a novel share-and-transfer framework to reach this goal with three
steps: (1) Convert each sentence in any language to language-
universal graph structures; in this paper we explore two approaches
based on universal dependency parses and complete graphs,
respectively. (2) Represent each node in the graph structure with a
cross-lingual word embedding so that all sentences in multiple
languages can be represented with one shared semantic space. (3)
Using this common semantic space, train event extractors from
English training data and apply them to languages that do not have
any event annotations. Experimental results on three languages
(Spanish, Russian and Ukrainian) without any annotations show this
framework achieves comparable performance to a state-of-the-art
supervised model trained from more than 1,500 manually annotated
event mentions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.243}
}
```

@InProceedings{ramponi-plank-lombardo:2020:LREC,
  author   = {Ramponi, Alan and Plank, Barbara and Lombardo, Rosario},
  title   = {Cross-Domain Evaluation of Edge Detection for Biomedical Event Extraction},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1982--1989},
  abstract = {Biomedical event extraction is a crucial task in order to automatically extract information from the increasingly growing body of biomedical literature. Despite advances in the methods in recent years, most event extraction systems are still evaluated in-domain and on complete event structures only. This makes it hard to determine the performance of intermediate stages of the task, such as edge detection, across different corpora. Motivated by these limitations, we present the first cross-domain study of edge detection for biomedical event extraction. We analyze differences between five existing gold standard corpora, create a standardized benchmark corpus, and provide a strong baseline model for edge detection. Experiments show a large drop in performance when the baseline is applied on out-of-domain data, confirming the need for domain adaptation methods for the task. To encourage research efforts in this direction, we make both the data and the baseline available to the research community: https://www.cosbi.eu/cfx/9985.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.244}
}

@InProceedings{thompson-EtAl:2020:LREC,
  author   = {Thompson, Paul and Yates, Tim and Inan, Emrah and Ananiadou, Sophia},
  title   = {Semantic Annotation for Improved Safety in Construction Work},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1990--1999},
  abstract = {Risk management is a vital activity to ensure employee safety in construction projects. Various documents provide important supporting evidence, including details of previous incidents, consequences and mitigation strategies. Potential hazards may depend on a complex set of project-specific attributes, including activities undertaken, location, equipment used, etc. However, finding evidence about previous projects with similar attributes can be problematic, since information about risks and mitigations is usually hidden within and may be dispersed across a

range of different free text documents. Automatic named entity recognition (NER), which identifies mentions of concepts in free text documents, is the first stage in structuring knowledge contained within them. While developing NER methods generally relies on annotated corpora, we are not aware of any such corpus targeted at concepts relevant to construction safety. In response, we have designed a novel named entity annotation scheme and associated guidelines for this domain, which covers hazards, consequences, mitigation strategies and project attributes. Four health and safety experts used the guidelines to annotate a total of 600 sentences from accident reports; an average inter-annotator agreement rate of 0.79 F-Score shows that our work constitutes an important first step towards developing tools for detailed semantic analysis of construction safety documents.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.245}
}

@InProceedings{tsekouras-EtAl:2020:LREC,
  author   = {Tsekouras, Leonidas and Petasis, Georgios and Giannakopoulos, George and Kosmopoulos, Aris},
  title   = {Social Web Observatory: A Platform and Method for Gathering Knowledge on Entities from Different Textual Sources},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2000--2008},
  abstract = {Within this work we describe a framework for the collection and summarization of information from the Web in an entity-driven manner. The framework consists of a set of appropriate workflows and the Social Web Observatory platform, which implements those workflows, supporting them through a language analysis pipeline. The pipeline includes text collection/crawling, identification of different entities, clustering of texts into events related to entities, entity-centric sentiment analysis, but also text analytics and visualization functionalities. The latter allow the user to take advantage of the gathered information as actionable knowledge: to understand the dynamics of the public opinion for a given entity over time and across real-world events. We describe the platform and the analysis functionality and evaluate the performance of the system, by allowing human users to score how the system fares in its intended purpose of summarizing entity-centered information from different sources in the Web.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.246}
}

@InProceedings{chaturvedi-EtAl:2020:LREC,
  author   = {Chaturvedi, Jaya and Viani, Natalia and Sanyal, Jyoti and Tytherleigh, Chloe and Hasan, Idil and Baird, Kate and Velupillai, Sumithra and Stewart, Robert and Roberts, Angus},
  title   = {Development of a Corpus Annotated with Medications

and their Attributes in Psychiatric Health Records},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2009--2016},
  abstract = {Free text fields within electronic health records (EHRs) contain valuable clinical information which is often missed when conducting research using EHR databases. One such type of information is medications which are not always available in structured fields, especially in mental health records. Most use cases that require medication information also generally require the associated temporal information (e.g. current or past) and attributes (e.g. dose, route, frequency). The purpose of this study is to develop a corpus of medication annotations in mental health records. The aim is to provide a more complete picture behind the mention of medications in the health records, by including additional contextual information around them, and to create a resource for use when developing and evaluating applications for the extraction of medications from EHR text. Thus far, an analysis of temporal information related to medications mentioned in a sample of mental health records has been conducted. The purpose of this analysis was to understand the complexity of medication mentions and their associated temporal information in the free text of EHRs, with a specific focus on the mental health domain.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.247}
}

@InProceedings{mandya-EtAl:2020:LREC,
  author   = {Mandya, Angrosh and O' Neill, James and Bollegala, Danushka and Coenen, Frans},
  title   = {Do not let the history haunt you: Mitigating Compounding Errors in Conversational Question Answering},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2017--2025},
  abstract = {The Conversational Question Answering (CoQA) task involves answering a sequence of inter-related conversational questions about a contextual paragraph. Although existing approaches employ human-written ground-truth answers for answering conversational questions at test time, in a realistic scenario, the CoQA model will not have any access to ground-truth answers for the previous questions, compelling the model to rely upon its own previously predicted answers for answering the subsequent questions. In this paper, we find that compounding errors occur when using previously predicted answers at test time, significantly lowering the performance of CoQA systems. To solve this problem, we propose a sampling strategy that dynamically selects between target answers

and model predictions during training, thereby closely simulating the situation at test time. Further, we analyse the severity of this phenomena as a function of the question type, conversation length and domain type.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.248}
}

@InProceedings{zeng-EtAl:2020:LREC,
  author   = {Zeng, Weixin and Zhao, Xiang and Tang, Jiuyang and Tan, Zhen and Huang, Xuqian},
  title   = {CLEEK: A Chinese Long-text Corpus for Entity Linking},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2026--2035},
  abstract = {Entity linking, as one of the fundamental tasks in natural language processing, is crucial to knowledge fusion, knowledge base construction and update. Nevertheless, in contrast to the research on entity linking for English text, which undergoes continuous development, the Chinese counterpart is still in its infancy. One prominent issue lies in publicly available annotated datasets and evaluation benchmarks, which are lacking and deficient. In specific, existing Chinese corpora for entity linking were mainly constructed from noisy short texts, such as microblogs and news headings, where long texts were largely overlooked, which yet constitute a wider spectrum of real-life scenarios. To address the issue, in this work, we build CLEEK, a Chinese corpus of multi-domain long text for entity linking, in order to encourage advancement of entity linking in languages besides English. The corpus consists of 100 documents from diverse domains, and is publicly accessible. Moreover, we devise a measure to evaluate the difficulty of documents with respect to entity linking, which is then used to characterize the corpus. Additionally, the results of two baselines and seven state-of-the-art solutions on CLEEK are reported and compared. The empirical results validate the usefulness of CLEEK and the effectiveness of proposed difficulty measure.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.249}
}

@InProceedings{smirnova-EtAl:2020:LREC,
  author   = {Smirnova, Katerina and Korotaev, Nikolay and Panikratova, Yana and Lebedeva, Irina and Pechenkova, Ekaterina and Fedorova, Olga},
  title   = {Using the RUPEX Multichannel Corpus in a Pilot fMRI Study on Speech Disfluencies},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},

```
  publisher      = {European Language Resources Association},
  pages     = {195--203},
  abstract  = {In modern linguistics and psycholinguistics speech disfluencies in real fluent speech are a well-known phenomenon. But it’s not still clear which components of brain systems are involved into its comprehension in a listener’s brain. In this paper we provide a pilot neuroimaging study of the possible neural correlates of speech disfluencies perception, using a combination of the corpus and functional magnetic-resonance imaging (fMRI) methods. Special technical procedure of selecting stimulus material from Russian multichannel corpus RUPEX allowed to create fragments in terms of requirements for the fMRI BOLD temporal resolution. They contain isolated speech disfluencies and their clusters. Also, we used the referential task for participants fMRI scanning. As a result, it was demonstrated that annotated multichannel corpora like RUPEX can be an important resource for experimental research in interdisciplinary fields. Thus, different aspects of communication can be explored through the prism of brain activation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.25}
}

@InProceedings{shafran-EtAl:2020:LREC,
  author    = {Shafran, Izhak  and  Du, Nan  and  Tran, Linh  and  Perry, Amanda  and  Keyes, Lauren  and  Knichel, Mark  and  Domin, Ashley  and  Huang, Lei  and  Chen, Yu-hui  and  Li, Gang  and  Wang, Mingqiu  and  El Shafey, Laurent  and  Soltau, Hagen  and  Paul, Justin Stuart},
  title     = {The Medical Scribe: Corpus Development and Model Performance Analyses},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2036--2044},
  abstract  = {There is a growing interest in creating tools to assist in clinical note generation using the audio of provider-patient encounters. Motivated by this goal and with the help of providers and medical scribes, we developed an annotation scheme to extract relevant clinical concepts. We used this annotation scheme to label a corpus of about 6k clinical encounters. This was used to train a state-of-the-art tagging model. We report ontologies, labeling results, model performances, and detailed analyses of the results. Our results show that the entities related to medications can be extracted with a relatively high accuracy of 0.90 F-score, followed by symptoms at 0.72 F-score, and conditions at 0.57 F-score. In our task, we not only identify where the symptoms are mentioned but also map them to canonical forms as they appear in the clinical notes. Of the different types of errors, in about 19-38\% of the cases, we find that the model output was correct, and about 17-32\% of the errors do not impact the clinical note. Taken together, the models developed in this work are more useful than the F-scores reflect, making it a promising approach for practical
```

```
applications.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.250}
}

@InProceedings{funaki-EtAl:2020:LREC,
  author    = {Funaki, Ruka  and  Nagata, Yusuke  and  Suenaga, Kohei  and  Mori, Shinsuke},
  title     = {A Contract Corpus for Recognizing Rights and Obligations},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2045--2053},
  abstract  = {A contract is a legal document executed by two or more parties. It is important for these parties to precisely understand their rights and obligations that are described in the contract. However, understanding the content of a contract is sometimes difficult and costly, particularly if the contract is long and complicated. Therefore, a language-processing system that can present information concerning rights and obligations found within a given contract document would help a contracting party to make better decisions. As a step toward the development of such a language-processing system, in this paper, we describe the annotated corpus of contract documents that we built. Our corpus is annotated so that a language-processing system can recognize a party's rights and obligations. The annotated information includes the parties involved in the contract, the rights and obligations of the parties, the conditions and the exceptions under which these rights and obligations to take effect. The corpus was built based on 46 English contracts and 25 Japanese contracts drafted by lawyers. We explain how we annotated the corpus and the statistics of the corpus. We also report the results of the experiments for recognizing rights and obligations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.251}
}

@InProceedings{pezanowski-mitra:2020:LREC,
  author    = {Pezanowski, Scott  and  Mitra, Prasenjit},
  title     = {Recognition of Implicit Geographic Movement in Text},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2054--2063},
  abstract  = {Analyzing the geographic movement of humans, animals, and other phenomena is a growing field of research. This research has benefited urban planning, logistics, animal migration understanding, and much more. Typically, the movement is captured as precise geographic coordinates and time stamps with Global
```

Positioning Systems (GPS). Although some research uses computational techniques to take advantage of implicit movement in descriptions of route directions, hiking paths, and historical exploration routes, innovation would accelerate with a large and diverse corpus. We created a corpus of sentences labeled as describing geographic movement or not and including the type of entity moving. Creating this corpus proved difficult without any comparable corpora to start with, high human labeling costs, and since movement can at times be interpreted differently. To overcome these challenges, we developed an iterative process employing hand labeling, crowd voting for confirmation, and machine learning to predict more labels. By merging advances in word embeddings with traditional machine learning models and model ensembling, prediction accuracy is at an acceptable level to produce a large silver-standard corpus despite the small gold-standard corpus training set. Our corpus will likely benefit computational processing of geography in text and spatial cognition, in addition to detection of movement.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.252}
}

@InProceedings{takamaru-EtAl:2020:LREC,
  author   = {Takamaru, Keiichi and Kimura, Yasutomo and Shibuki, Hideyuki and Ototake, Hokuto and Uchida, Yuzu and Sakamoto, Kotaro and Ishioroshi, Madoka and Mitamura, Teruko and Kando, Noriko},
  title   = {Extraction of the Argument Structure of Tokyo Metropolitan Assembly Minutes: Segmentation of Question-and-Answer Sets},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2064--2068},
  abstract = {In this study, we construct a corpus of Japanese local assembly minutes. All speeches in an assembly were transcribed into a local assembly minutes based on the local autonomy law. Therefore, the local assembly minutes form an extremely large amount of text data. Our ultimate objectives were to summarize and present the arguments in the assemblies, and to use the minutes as primary information for arguments in local politics. To achieve this, we structured all statements in assembly minutes. We focused on the structure of the discussion, i.e., the extraction of question and answer pairs. We organized the shared task ``QA Lab-PoliInfo'' in NTCIR 14. We conducted a ``segmentation task'' to identify the scope of one question and answer in the minutes as a sub task of the shared task. For the segmentation task, 24 runs from five teams were submitted. Based on the obtained results, the best recall was 1.000, best precision was 0.940, and best F-measure was 0.895.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.253}
}

@InProceedings{robin-EtAl:2020:LREC,

```
  author    = {Robin, Cécile  and  Isazad Mashinchi, Mona  and
Ahmadi Zeleti, Fatemeh  and  Ojo, Adegboyega  and  Buitelaar, Paul},
  title     = {A Term Extraction Approach to Survey Analysis in
Health Care},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2069--2077},
  abstract  = {The voice of the customer has for a long time been a
key focus of businesses in all domains. It has received a lot of
attention from the research community in Natural Language Processing
(NLP) resulting in many approaches to analyzing customers feedback
((aspect-based) sentiment analysis, topic modeling, etc.). In the
health domain, public and private bodies are increasingly
prioritizing patient engagement for assessing the quality of the
service given at each stage of the care. Patient and customer
satisfaction analysis relate in many ways. In the domain of health
particularly, a more precise and insightful analysis is needed to
help practitioners locate potential issues and plan actions
accordingly. We introduce here an approach to patient experience
with the analysis of free text questions from the 2017 Irish
National Inpatient Survey campaign using term extraction as a means
to highlight important and insightful subject matters raised by
patients. We evaluate the results by mapping them to a manually
constructed framework following the Activity, Resource, Context
(ARC) methodology (Ordenes, 2014) and specific to the health care
environment, and compare our results against manual annotations done
on the full 2017 dataset based on those categories.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.254}
}

@InProceedings{kruiper-EtAl:2020:LREC,
  author    = {Kruiper, Ruben  and  Vincent, Julian F.V.  and  Chen-
Burger, Jessica  and  Desmulliez, Marc P.Y.  and  Konstas, Ioannis},
  title     = {A Scientific Information Extraction Dataset for
Nature Inspired Engineering},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2078--2085},
  abstract  = {Nature has inspired various ground-breaking
technological developments in applications ranging from robotics to
aerospace engineering and the manufacturing of medical devices.
However, accessing the information captured in scientific biology
texts is a time-consuming and hard task that requires domain-
specific knowledge. Improving access for outsiders can help
interdisciplinary research like Nature Inspired Engineering. This
paper describes a dataset of 1,500 manually-annotated sentences that
```

express domain-independent relations between central concepts in a scientific biology text, such as trade-offs and correlations. The arguments of these relations can be Multi Word Expressions and have been annotated with modifying phrases to form non-projective graphs. The dataset allows for training and evaluating Relation Extraction algorithms that aim for coarse-grained typing of scientific biological documents, enabling a high-level filter for engineers.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.255}
}

@InProceedings{vanetik-EtAl:2020:LREC,
  author   = {Vanetik, Natalia and Litvak, Marina and Shevchuk, Sergey and Reznik, Lior},
  title   = {Automated Discovery of Mathematical Definitions in Text},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2086--2094},
  abstract = {Automatic definition extraction from texts is an important task that has numerous applications in several natural language processing fields such as summarization, analysis of scientific texts, automatic taxonomy generation, ontology generation, concept identification, and question answering. For definitions that are contained within a single sentence, this problem can be viewed as a binary classification of sentences into definitions and non-definitions. Definitions in scientific literature can be generic (Wikipedia) or more formal (mathematical articles). In this paper, we focus on automatic detection of one-sentence definitions in mathematical texts, which are difficult to separate from surrounding text. We experiment with several data representations, which include sentence syntactic structure and word embeddings, and apply deep learning methods such as convolutional neural network (CNN) and recurrent neural network (RNN), in order to identify mathematical definitions. Our experiments demonstrate the superiority of CNN and its combination with RNN, applied on the syntactically-enriched input representation. We also present a new dataset for definition extraction from mathematical texts. We demonstrate that the use of this dataset for training learning models improves the quality of definition extraction when these models are then used for other definition datasets. Our experiments with different domains approve that mathematical definitions require special treatment, and that using cross-domain learning is inefficient.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.256}
}

@InProceedings{wu-EtAl:2020:LREC2,
  author   = {Wu, Chuan and Kanoulas, Evangelos and de Rijke, Maarten and Lu, Wei},
  title   = {WN-Salience: A Corpus of News Articles with Entity

Salience Annotations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2095--2102},
  abstract = {Entities can be found in various text genres, ranging from tweets and web pages to user queries submitted to web search engines. Existing research either considers all entities in the text equally important, or heuristics are used to measure their salience. We believe that a key reason for the relatively limited work on entity salience is the lack of appropriate datasets. To support research on entity salience, we present a new dataset, the WikiNews Salience dataset (WN-Salience), which can be used to benchmark tasks such as entity salience detection and salient entity linking. WN-Salience is built on top of Wikinews, a Wikimedia project whose mission is to present reliable news articles. Entities in Wikinews articles are identified by the authors of the articles and are linked to Wikinews categories when they are salient or to Wikipedia pages otherwise. The dataset is built automatically, and consists of approximately 7,000 news articles, and 90,000 in-text entity annotations. We compare the WN-Salience dataset against existing datasets on the task and analyze their differences. Furthermore, we conduct experiments on entity salience detection; the results demonstrate that WN-Salience is a challenging testbed that is complementary to existing ones.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.257}
}

@InProceedings{tadesse-tsegaye-qaqqabaa:2020:LREC,
  author   = {tadesse, ephrem and Tsegaye, Rosa and Qaqqabaa, Kuulaa},
  title   = {Event Extraction from Unstructured Amharic Text},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2103--2109},
  abstract = {In information extraction, event extraction is one of the types that extract the specific knowledge of certain incidents from texts. Event extraction has been done on different languages text but not on one of the Semitic language, Amharic. In this study, we present a system that extracts an event from unstructured Amharic text. The system has designed by the integration of supervised machine learning and rule-based approaches. We call this system a hybrid system. The system uses the supervised machine learning to detect events from the text and the handcrafted and the rule-based rules to extract the event from the text. For the event extraction, we have been using event arguments. Event arguments identify event triggering words or phrases that clearly express the occurrence of

the event. The event argument attributes can be verbs, nouns, sometimes adjectives (such as˜rg/wedding) and time as well. The hybrid system has compared with the standalone rule-based method that is well known for event extraction. The study has shown that the hybrid system has outperformed the standalone rule-based method.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.258}
}

@InProceedings{magnini-lavelli-magnolini:2020:LREC,
  author   = {Magnini, Bernardo and Lavelli, Alberto and Magnolini, Simone},
  title   = {Comparing Machine Learning and Deep Learning Approaches on NLP Tasks for the Italian Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2110--2119},
  abstract = {We present a comparison between deep learning and traditional machine learning methods for various NLP tasks in Italian. We carried on experiments using available datasets (e.g., from the Evalita shared tasks) on two sequence tagging tasks (i.e., named entities recognition and nominal entities recognition) and four classification tasks (i.e., lexical relations among words, semantic relations among sentences, sentiment analysis and text classification). We show that deep learning approaches outperform traditional machine learning algorithms in sequence tagging, while for classification tasks that heavily rely on semantics approaches based on feature engineering are still competitive. We think that a similar analysis could be carried out for other languages to provide an assessment of machine learning / deep learning models across different languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.259}
}

@InProceedings{koyama-EtAl:2020:LREC,
  author   = {Koyama, Aomi and Kiyuna, Tomoshige and Kobayashi, Kenji and Arai, Mio and Komachi, Mamoru},
  title   = {Construction of an Evaluation Corpus for Grammatical Error Correction for Learners of Japanese as a Second Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {204--211},
  abstract = {The NAIST Lang-8 Learner Corpora (Lang-8 corpus) is one of the largest second-language learner corpora. The Lang-8 corpus is suitable as a training dataset for machine translation-based grammatical error correction systems. However, it is not

```
suitable as an evaluation dataset because the corrected sentences
sometimes include inappropriate sentences. Therefore, we created and
released an evaluation corpus for correcting grammatical errors made
by learners of Japanese as a Second Language (JSL). As our corpus
has less noise and its annotation scheme reflects the
characteristics of the dataset, it is ideal as an evaluation corpus
for correcting grammatical errors in sentences written by JSL
learners. In addition, we applied neural machine translation (NMT)
and statistical machine translation (SMT) techniques to correct the
grammar of the JSL learners' sentences and evaluated their results
using our corpus. We also compared the performance of the NMT system
with that of the SMT system.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.26}
}

@InProceedings{nabizadeh-kolossa-heckmann:2020:LREC,
  author    = {Nabizadeh, Nima  and  Kolossa, Dorothea  and
Heckmann, Martin},
  title     = {MyFixit: An Annotated Dataset, Annotation Tool, and
Baseline Methods for Information Extraction from Repair Manuals},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2120--2128},
  abstract  = {Text instructions are among the most widely used
media for learning and teaching. Hence, to create assistance systems
that are capable of supporting humans autonomously in new tasks, it
would be immensely productive, if machines were enabled to extract
task knowledge from such text instructions. In this paper, we,
therefore, focus on information extraction (IE) from the
instructional text in repair manuals. This brings with it the
multiple challenges of information extraction from the situated and
technical language in relatively long and often complex
instructions. To tackle these challenges, we introduce a semi-
structured dataset of repair manuals. The dataset is annotated in a
large category of devices, with information that we consider most
valuable for an automated repair assistant, including the required
tools and the disassembled parts at each step of the repair
progress. We then propose methods that can serve as baselines for
this IE task: an unsupervised method based on a bags-of-n-grams
similarity for extracting the needed tools in each repair step, and
a deep-learning-based sequence labeling model for extracting the
identity of disassembled parts. These baseline methods are
integrated into a semi-automatic web-based annotator application
that is also available along with the dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.260}
}

@InProceedings{vanerp-groth:2020:LREC,
  author    = {van Erp, Marieke  and  Groth, Paul},
  title     = {Towards Entity Spaces},
```

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2129--2137},
  abstract  = {Entities are a central element of knowledge bases and
are important input to many knowledge-centric tasks including text
analysis. For example, they allow us to find documents relevant to a
specific entity irrespective of the underlying syntactic expression
within a document. However, the entities that are commonly
represented in knowledge bases are often a simplification of what is
truly being referred to in text. For example, in a knowledge base,
we may have an entity for Germany as a country but not for the more
fuzzy concept of Germany that covers notions of German Population,
German Drivers, and the German Government. Inspired by recent
advances in contextual word embeddings, we introduce the concept of
entity spaces - specific representations of a set of associated
entities with near-identity. Thus, these entity spaces provide a
handle to an amorphous grouping of entities. We developed a proof-
of-concept for English showing how, through the introduction of
entity spaces in the form of disambiguation pages, the recall of
entity linking can be improved.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.261}
}

@InProceedings{fell-EtAl:2020:LREC,
  author    = {Fell, Michael  and  Cabrio, Elena  and  Korfed,
Elmahdi  and  Buffa, Michel  and  Gandon, Fabien},
  title     = {Love Me, Love Me, Say (and Write!) that You Love Me:
Enriching the WASABI Song Corpus with Lyrics Annotations},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2138--2147},
  abstract  = {We present the WASABI Song Corpus, a large corpus of
songs enriched with metadata extracted from music databases on the
Web, and resulting from the processing of song lyrics and from audio
analysis. More specifically, given that lyrics encode an important
part of the semantics of a song, we focus here on the description of
the methods we proposed to extract relevant information from the
lyrics, as their structure segmentation, their topic, the
explicitness of the lyrics content, the salient passages of a song
and the emotions conveyed. The creation of the resource is still
ongoing: so far, the corpus contains 1.73M songs with lyrics (1.41M
unique lyrics) annotated at different levels with the output of the
above mentioned methods. Such corpus labels and the provided methods
can be exploited by music search engines and music professionals
(e.g. journalists, radio presenters) to better handle large
collections of lyrics, allowing an intelligent browsing,
```

```
categorization and segmentation recommendation of songs.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.262}
}

@InProceedings{ocal-finlayson:2020:LREC,
  author    = {Ocal, Mustafa  and  Finlayson, Mark},
  title     = {Evaluating Information Loss in Temporal Dependency
Trees},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2148--2156},
  abstract  = {Temporal Dependency Trees (TDTs) have emerged as an
alternative to full temporal graphs for representing the temporal
structure of texts, with a key advantage being that TDTs can be
straightforwardly computed using adapted dependency parsers.
Relative to temporal graphs, the tree form of TDTs naturally omits
some fraction of temporal relationships, which intuitively should
decrease the amount of temporal information available, potentially
increasing temporal indeterminacy of the global ordering. We
demonstrate a new method for quantifying this indeterminacy that
relies on solving temporal constraint problems to extract timelines,
and show that TDTs result in up to a 109\% increase in temporal
indeterminacy over their corresponding temporal graphs for the three
corpora we examine. On average, the increase in indeterminacy is 32\
%, and we show that this increase is a result of the TDT
representation eliminating on average only 2.4\% of total temporal
relations. This result suggests that small differences can have big
effects in temporal graphs, and the use of TDTs must be balanced
against their deficiencies, with tasks requiring an accurate global
temporal ordering potentially calling for use of the full temporal
graph},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.263}
}

@InProceedings{humphreys-EtAl:2020:LREC,
  author    = {Humphreys, Llio  and  Boella, Guido  and  Di Caro,
Luigi  and  Robaldo, Livio  and  van der Torre, Leon  and
Ghanavati, Sepideh  and  Muthuri, Robert},
  title     = {Populating Legal Ontologies using Semantic Role
Labeling},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2157--2166},
  abstract  = {This paper is concerned with the goal of maintaining
legal information and compliance systems: the ‘resource consumption
bottleneck’ of creating semantic technologies manually. The use of
```

automated information extraction techniques could significantly reduce this bottleneck. The research question of this paper is: How to address the resource bottleneck problem of creating specialist knowledge management systems? In particular, how to semi-automate the extraction of norms and their elements to populate legal ontologies? This paper shows that the acquisition paradox can be addressed by combining state-of-the-art general-purpose NLP modules with pre- and post-processing using rules based on domain knowledge. It describes a Semantic Role Labeling based information extraction system to extract norms from legislation and represent them as structured norms in legal ontologies. The output is intended to help make laws more accessible, understandable, and searchable in legal document management systems such as Eunomos (Boella et al., 2016).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.264}
}

@InProceedings{marciczuk-oleksy-wieczorek:2020:LREC,
  author   = {Marcińczuk, Michał and Oleksy, Marcin and Wieczorek, Jan},
  title   = {PST 2.0 – Corpus of Polish Spatial Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2167--2174},
  abstract = {In the paper, we focus on modeling spatial expressions in texts. We present the guidelines used to annotate the PST 2.0 (Corpus of Polish Spatial Texts) – a corpus designed for training and testing the tools for spatial expression recognition. The corpus contains a set of texts gathered from texts collected from travel blogs available under Creative Commons license. We have defined our guidelines based on three existing specifications for English (SpatialML, SpatialRole Labelling from SemEval-2013 Task 3 and ISO-Space1.4 from SpaceEval 2014). We briefly present the existing specifications and discuss what modifications have been made to adapt the guidelines to the characteristics of the Polish language. We also describe the process of data collection and manual annotation, including inter-annotator agreement calculation and corpus statistics. In the end, we present detailed statistics of the PST 2.0 corpus, which include the number of components, relations, expressions, and the most common values of spatial indicators, motion indicators, path indicators, distances, directions, and regions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.265}
}

@InProceedings{ferreira-freitas:2020:LREC,
  author   = {Ferreira, Deborah and Freitas, André},
  title   = {Natural Language Premise Selection: Finding Supporting Statements for Mathematical Text},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2175--2182},
  abstract  = {Mathematical text is written using a combination of
words and mathematical expressions. This combination, along with a
specific way of structuring sentences makes it challenging for
state-of-art NLP tools to understand and reason on top of
mathematical discourse. In this work, we propose a new NLP task, the
natural premise selection, which is used to retrieve supporting
definitions and supporting propositions that are useful for
generating an informal mathematical proof for a particular
statement. We also make available a dataset, NL-PS, which can be
used to evaluate different approaches for the natural premise
selection task. Using different baselines, we demonstrate the
underlying interpretation challenges associated with the task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.266}
}

@InProceedings{valenzuelaescrcega-hahnpowell-bell:2020:LREC,
  author    = {Valenzuela-Escárcega, Marco A.  and  Hahn-Powell, Gus
and  Bell, Dane},
  title     = {Odinson: A Fast Rule-based Information Extraction
Framework},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2183--2191},
  abstract  = {We present Odinson, a rule-based information
extraction framework, which couples a simple yet powerful pattern
language that can operate over multiple representations of text,
with a runtime system that operates in near real time. In the
Odinson query language, a single pattern may combine regular
expressions over surface tokens with regular expressions over graphs
such as syntactic dependencies. To guarantee the rapid matching of
these patterns, our framework indexes most of the necessary
information for matching patterns, including directed graphs such as
syntactic dependencies, into a custom Lucene index. Indexing
minimizes the amount of expensive pattern matching that must take
place at runtime. As a result, the runtime system matches a syntax-
based graph traversal in 2.8 seconds in a corpus of over 134 million
sentences, nearly 150,000 times faster than its predecessor.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.267}
}

@InProceedings{dsouza-EtAl:2020:LREC,
  author    = {D'Souza, Jennifer  and  Hoppe, Anett  and  Brack,
Arthur  and  Jaradeh, Mohmad Yaser  and  Auer, Sören  and  Ewerth,
Ralph},
  title     = {The STEM-ECR Dataset: Grounding Scientific Entity
```

```
References in STEM Scholarly Content to Authoritative Encyclopedic
and Lexicographic Sources},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2192--2203},
  abstract  = {We introduce the STEM (Science, Technology,
Engineering, and Medicine) Dataset for Scientific Entity Extraction,
Classification, and Resolution, version 1.0 (STEM-ECR v1.0). The
STEM-ECR v1.0 dataset has been developed to provide a benchmark for
the evaluation of scientific entity extraction, classification, and
resolution tasks in a domain-independent fashion. It comprises
abstracts in 10 STEM disciplines that were found to be the most
prolific ones on a major publishing platform. We describe the
creation of such a multidisciplinary corpus and highlight the
obtained findings in terms of the following features: 1) a generic
conceptual formalism for scientific entities in a multidisciplinary
scientific context; 2) the feasibility of the domain-independent
human annotation of scientific entities under such a generic
formalism; 3) a performance benchmark obtainable for automatic
extraction of multidisciplinary scientific entities using BERT-based
neural models; 4) a delineated 3-step entity resolution procedure
for human annotation of the scientific entities via encyclopedic
entity linking and lexicographic word sense disambiguation; and 5)
human evaluations of Babelfy returned encyclopedic links and
lexicographic senses for our entities. Our findings cumulatively
indicate that human annotation and automatic learning of
multidisciplinary scientific concepts as well as their semantic
disambiguation in a wide-ranging setting as STEM is reasonable.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.268}
}

@InProceedings{alexeeva-EtAl:2020:LREC,
  author    = {Alexeeva, Maria  and  Sharp, Rebecca  and
Valenzuela-Escárcega, Marco A.  and  Kadowaki, Jennifer  and
Pyarelal, Adarsh  and  Morrison, Clayton},
  title     = {MathAlign: Linking Formula Identifiers to their
Contextual Natural Language Descriptions},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2204--2212},
  abstract  = {Extending machine reading approaches to extract
mathematical concepts and their descriptions is useful for a variety
of tasks, ranging from mathematical information retrieval to
increasing accessibility of scientific documents for the visually
impaired. This entails segmenting mathematical formulae into
identifiers and linking them to their natural language descriptions.
```

```
We propose a rule-based approach for this task, which extracts LaTeX
representations of formula identifiers and links them to their in-
text descriptions, given only the original PDF and the location of
the formula of interest. We also present a novel evaluation dataset
for this task, as well as the tool used to create it.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.269}
}

@InProceedings{nam-EtAl:2020:LREC,
  author    = {Nam, Sangha  and  Lee, Minho  and  Kim, Donghwan  and
Han, Kijong  and  Kim, Kuntae  and  Yoon, Sooji  and  Kim, Eun-kyung
and  Choi, Key-Sun},
  title     = {Effective Crowdsourcing of Multiple Tasks for
Comprehensive Knowledge Extraction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {212--219},
  abstract  = {Information extraction from unstructured texts plays
a vital role in the field of natural language processing. Although
there has been extensive research into each information extraction
task (i.e., entity linking, coreference resolution, and relation
extraction), data are not available for a continuous and coherent
evaluation of all information extraction tasks in a comprehensive
framework. Given that each task is performed and evaluated with a
different dataset, analyzing the effect of the previous task on the
next task with a single dataset throughout the information
extraction process is impossible. This paper aims to propose a
Korean information extraction initiative point and promote research
in this field by presenting crowdsourcing data collected for four
information extraction tasks from the same corpus and the training
and evaluation results for each task of a state-of-the-art model.
These machine learning data for Korean information extraction are
the first of their kind, and there are plans to continuously
increase the data volume. The test results will serve as an
initiative result for each Korean information extraction task and
are expected to serve as a comparison target for various studies on
Korean information extraction using the data collected in this
study.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.27}
}

@InProceedings{sainz-EtAl:2020:LREC,
  author    = {Sainz, Oscar  and  Lopez de Lacalle, Oier  and
Aldabe, Itziar  and  Maritxalar, Montse},
  title     = {Domain Adapted Distant Supervision for Pedagogically
Motivated Relation Extraction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2213--2222},
  abstract  = {In this paper we present a relation extraction system
that given a text extracts pedagogically motivated relation types,
as a previous step to obtaining a semantic representation of the
text which will make possible to automatically generate questions
for reading comprehension. The system maps pedagogically motivated
relations with relations from ConceptNet and deploys Distant
Supervision for relation extraction. We run a study on a subset of
those relationships in order to analyse the viability of our
approach. For that, we build a domain-specific relation extraction
system and explore two relation extraction models: a state-of-the-
art model based on transfer learning and a discrete feature based
machine learning model. Experiments show that the neural model
obtains better results in terms of F-score and we yield promising
results on the subset of relations suitable for pedagogical
purposes. We thus consider that distant supervision for relation
extraction is a valid approach in our target domain, i.e. biology.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.270}
}

@InProceedings{niu-EtAl:2020:LREC,
  author    = {Niu, Jingcheng  and  Ng, Victoria  and  Penn, Gerald
and  Rees, Erin E.},
  title     = {Temporal Histories of Epidemic Events (THEE): A Case
Study in Temporal Annotation for Public Health},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2223--2230},
  abstract  = {We present a new temporal annotation standard, THEE-
TimeML, and a corpus TheeBank enabling precise temporal information
extraction (TIE) for event-based surveillance (EBS) systems in the
public health domain. Current EBS must estimate the occurrence time
of each event based on coarse document metadata such as document
publication time. Because of the complicated language and narration
style of news articles, estimated case outbreak times are often
inaccurate or even erroneous. Thus, it is necessary to create
annotation standards and corpora to facilitate the development of
TIE systems in the public health domain to address this problem.We
will discuss the adaptations that have proved necessary for this
domain as we present THEE-TimeML and TheeBank. Finally, we document
the corpus annotation process, and demonstrate the immediate benefit
to public health applications brought by the annotations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.271}
}

@InProceedings{khadka-cantador-fernandez:2020:LREC,
  author    = {Khadka, Anita  and  Cantador, Iván  and  Fernandez,
Miriam},
```

```
  title     = {Exploiting Citation Knowledge in Personalised
Recommendation of Recent Scientific Publications},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2231--2240},
  abstract  = {In this paper we address the problem of providing
personalised recommendations of recent scientific publications to a
particular user, and explore the use of citation knowledge to do so.
For this purpose, we have generated a novel dataset that captures
authors' publication history and is enriched with different forms of
paper citation knowledge, namely citation graphs, citation
positions, citation contexts, and citation types. Through a number
of empirical experiments on such dataset, we show that the
exploitation of the extracted knowledge, particularly the type of
citation, is a promising approach for recommending recently
published papers that may not be cited yet. The dataset, which we
make publicly available, also represents a valuable resource for
further investigation on academic information retrieval and
filtering.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.272}
}

@InProceedings{sahoo-EtAl:2020:LREC,
  author    = {Sahoo, Sovan Kumar  and  Saha, Saumajit  and  Ekbal,
Asif  and  Bhattacharyya, Pushpak},
  title     = {A Platform for Event Extraction in Hindi},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2241--2250},
  abstract  = {Event Extraction is an important task in the
widespread field of Natural Language Processing (NLP). Though this
task is adequately addressed in English with sufficient resources,
we are unaware of any benchmark setup in Indian languages. Hindi is
one of the most widely spoken languages in the world. In this paper,
we present an Event Extraction framework for Hindi language by
creating an annotated resource for benchmarking, and then developing
deep learning based models to set as the baselines. We crawl more
than seventeen hundred disaster related Hindi news articles from the
various news sources. We also develop deep learning based models for
Event Trigger Detection and Classification, Argument Detection and
Classification and Event-Argument Linking.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.273}
}

@InProceedings{datta-EtAl:2020:LREC,
  author    = {Datta, Surabhi  and  Ulinski, Morgan  and  Godfrey-
```

Stovall, Jordan and Khanpara, Shekhar and Riascos-Castaneda, Roy F. and Roberts, Kirk},
  title   = {Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2251--2260},
  abstract = {This paper proposes a representation framework for encoding spatial language in radiology based on frame semantics. The framework is adopted from the existing SpatialNet representation in the general domain with the aim to generate more accurate representations of spatial language used by radiologists. We describe Rad-SpatialNet in detail along with illustrating the importance of incorporating domain knowledge in understanding the varied linguistic expressions involved in different radiological spatial relations. This work also constructs a corpus of 400 radiology reports of three examination types (chest X-rays, brain MRIs, and babygrams) annotated with fine-grained contextual information according to this schema. Spatial trigger expressions and elements corresponding to a spatial frame are annotated. We apply BERT-based models (BERT-Base and BERT- Large) to first extract the trigger terms (lexical units for a spatial frame) and then to identify the related frame elements. The results of BERT- Large are decent, with F1 of 77.89 for spatial trigger extraction and an overall F1 of 81.61 and 66.25 across all frame elements using gold and predicted spatial triggers respectively. This frame-based resource can be used to develop and evaluate more advanced natural language processing (NLP) methods for extracting fine-grained spatial information from radiology text in the future.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.274}
}

@InProceedings{masson-paroubek:2020:LREC,
  author   = {Masson, Corentin and Paroubek, Patrick},
  title   = {NLP Analytics in Finance with DoRe: A French 250M Tokens Corpus of Corporate Annual Reports},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2261--2267},
  abstract = {Recent advances in neural computing and word embeddings for semantic processing open many new applications areas which had been left unaddressed so far because of inadequate language understanding capacity. But this new kind of approaches rely even more on training data to be operational. Corpora for financial applications exists, but most of them concern stock market prediction and are in English. To address this need for the French

language and regulation oriented applications which require a deeper understanding of the text content, we hereby present “DoRe”, a French and dialectal French Corpus for NLP analytics in Finance, Regulation and Investment. This corpus is composed of: (a) 1769 Annual Reports from 336 companies among the most capitalized companies in: France (Euronext Paris) \& Belgium (Euronext Brussels), covering a time frame from 2009 to 2019, and (b) related MetaData containing information for each company about its ISIN code, capitalization and sector. This corpus is designed to be as modular as possible in order to allow for maximum reuse in different tasks pertaining to Economics, Finance and Regulation. After presenting existing resources, we relate the construction of the DoRe corpus and the rationale behind our choices, concluding on the spectrum of possible uses of this new resource for NLP applications.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.275}
}

@InProceedings{maldonado-harabagiu:2020:LREC,
  author   = {Maldonado, Ramon and Harabagiu, Sanda},
  title   = {The Language of Brain Signals: Natural Language Processing of Electroencephalography Reports},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2268--2275},
  abstract = {Brain signals are captured by clinical electroencephalography (EEG) which is an excellent tool for probing neural function. When EEG tests are performed, a textual EEG report is generated by the neurologist to document the findings, thus using language that describes the brain signals and its clinical correlations. Even with the impetus provided by the BRAIN initiative (brainititiative.nih.gov), there are no annotations available in texts that capture language describing the brain activities and their correlations with various pathologies. In this paper we describe an annotation effort carried out on a large corpus of EEG reports, providing examples of EEG-specific and clinically relevant concepts. In addition, we detail our annotation schema for brain signal attributes. We also discuss the resulting annotation of long-distance relations between concepts in EEG reports. By exemplifying a self-attention joint-learning to predict similar annotations in the EEG report corpus, we discuss the promising results, hoping that our effort will inform the design of novel knowledge capture techniques that will include the language of brain signals.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.276}
}

@InProceedings{shavrina-EtAl:2020:LREC,
  author   = {Shavrina, Tatiana and Emelyanov, Anton and Fenogenova, Alena and Fomin, Vadim and Mikhailov, Vladislav and Evlampiev, Andrey and Malykh, Valentin and Larin, Vladimir and

Natekin, Alex and Vatulin, Aleksandr and Romov, Peter and Anastasiev, Daniil and Zinov, Nikolai and Chertok, Andrey},
  title   = {Humans Keep It One Hundred: an Overview of AI Journey},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2276--2284},
  abstract = {Artificial General Intelligence (AGI) is showing growing performance in numerous applications – beating human performance in Chess and Go, using knowledge bases and text sources to answer questions (SQuAD) and even pass human examination (Aristo project). In this paper, we describe the results of AI Journey, a competition of AI-systems aimed to improve AI performance on knowledge bases, reasoning and text generation. Competing systems pass the final native language exam (in Russian), including versatile grammar tasks (test and open questions) and an essay, achieving a high score of 69\%, with 68\% being an average human result. During the competition, a baseline for the task and essay parts was proposed, and 80+ systems were submitted, showing different approaches to task understanding and reasoning. All the data and solutions can be found on github https://github.com/sberbank-ai/combined_solution_aij2019},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.277}
}

@InProceedings{deboer-verhoosel:2020:LREC,
  author   = {de Boer, Maaike and Verhoosel, Jack P. C.},
  title   = {Towards Data-driven Ontologies: a Filtering Approach using Keywords and Natural Language Constructs},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2285--2292},
  abstract = {Creating ontologies is an expensive task. Our vision is that we can automatically generate ontologies based on a set of relevant documents to create a kick-start in ontology creating sessions. In this paper, we focus on enhancing two often used methods, OpenIE and co-occurrences. We evaluate the methods on two document sets, one about pizza and one about the agriculture domain. The methods are evaluated using two types of F1-score (objective, quantitative) and through a human assessment (subjective, qualitative). The results show that 1) Cooc performs both objectively and subjectively better than OpenIE; 2) the filtering methods based on keywords and on Word2vec perform similarly; 3) the filtering methods both perform better compared to OpenIE and similar to Cooc; 4) Cooc-NVP performs best, especially considering the subjective evaluation. Although, the investigated methods provide a

good start for extracting an ontology out of a set of domain documents, various improvements are still possible, especially in the natural language based methods.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.278}
}

@InProceedings{jabbari-EtAl:2020:LREC,
  author   = {Jabbari, Ali and Sauvage, Olivier and Zeine, Hamada and Chergui, Hamza},
  title   = {A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2293--2299},
  abstract = {In financial services industry, compliance involves a series of practices and controls in order to meet key regulatory standards which aim to reduce financial risk and crime, e.g.\ money laundering and financing of terrorism. Faced with the growing risks, it is imperative for financial institutions to seek automated information extraction techniques for monitoring financial activities of their customers. This work describes an ontology of compliance-related concepts and relationships along with a corpus annotated according to it. The presented corpus consists of financial news articles in French and allows for training and evaluating domain-specific named entity recognition and relation extraction algorithms. We present some of our experimental results on named entity recognition and relation extraction using our annotated corpus. We aim to furthermore use the the proposed ontology towards construction of a knowledge base of financial relations.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.279}
}

@InProceedings{roque-EtAl:2020:LREC,
  author   = {Roque, Antonio and Tsuetaki, Alexander and Sarathy, Vasanth and Scheutz, Matthias},
  title   = {Developing a Corpus of Indirect Speech Act Schemas},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {220--228},
  abstract = {Resolving Indirect Speech Acts (ISAs), in which the intended meaning of an utterance is not identical to its literal meaning, is essential to enabling the participation of intelligent systems in peoples' everyday lives. Especially challenging are those cases in which the interpretation of such ISAs depends on context. To test a system's ability to perform ISA resolution we need a

```
corpus, but developing such a corpus is difficult, especialy given
the contex-dependent requirement. This paper addresses the difficult
problems of constructing a corpus of ISAs, taking inspiration from
relevant work in using corpora for reasoning tasks. We present a
formal representation of ISA Schemas required for such testing,
including a measure of the difficulty of a particular schema. We
develop an approach to authoring these schemas using corpus analysis
and crowdsourcing, to maximize realism and minimize the amount of
expert authoring needed. Finally, we describe several
characteristics of collected data, and potential future work.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.28}
}

@InProceedings{bebeshina-lafourcade:2020:LREC,
  author    = {Bebeshina, Nadia  and  Lafourcade, Mathieu},
  title     = {Inferences for Lexical Semantic Resource Building
with Less Supervision},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2300--2305},
  abstract  = {Lexical semantic resources may be built using various
approaches such as extraction from corpora, integration of the
relevant pieces of knowledge from the pre-existing knowledge
resources, and endogenous inference. Each of these techniques needs
human supervision in order to deal with the potential errors,
mapping difficulties or inferred candidate validation. We detail how
various inference processes can be employed for the less supervised
lexical semantic resource building. Our experience is based on the
combination of different inference techniques for multilingual
resource building and evaluation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.280}
}

@InProceedings{iwai-EtAl:2020:LREC1,
  author    = {Iwai, Ritsuko  and  Kawahara, Daisuke  and  Kumada,
Takatsune  and  Kurohashi, Sadao},
  title     = {Acquiring Social Knowledge about Personality and
Driving-related Behavior},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2306--2315},
  abstract  = {In this paper, we introduce our psychological
approach to collect human-specific social knowledge from a text
corpus, using NLP techniques. It is often not explicitly described
but shared among people, which we call social knowledge. We focus on
the social knowledge, especially personality and driving. We used
```

the language resources that were developed based on psychological research methods; a Japanese personality dictionary (317 words) and a driving experience corpus (8,080 sentences) annotated with behavior and subjectivity. Using them, we automatically extracted collocations between personality descriptors and driving-related behavior from a driving behavior and subjectivity corpus (1,803,328 sentences after filtering) and obtained unique 5,334 collocations. To evaluate the collocations as social knowledge, we designed four step-by-step crowdsourcing tasks. They resulted in 266 pieces of social knowledge. They include the knowledge that might be difficult to recall by themselves but easy to agree with. We discuss the acquired social knowledge and the contribution to implementations into systems.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.281}
}

@InProceedings{becker-korfhage-frank:2020:LREC,
  author   = {Becker, Maria and Korfhage, Katharina and Frank, Anette},
  title   = {Implicit Knowledge in Argumentative Texts: An Annotated Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2316--2324},
  abstract = {When speaking or writing, people omit information that seems clear and evident, such that only part of the message is expressed in words. Especially in argumentative texts it is very common that (important) parts of the argument are implied and omitted. We hypothesize that for argument analysis it will be beneficial to reconstruct this implied information. As a starting point for filling knowledge gaps, we build a corpus consisting of high-quality human annotations of missing and implied information in argumentative texts. To learn more about the characteristics of both the argumentative texts and the added information, we further annotate the data with semantic clause types and commonsense knowledge relations. The outcome of our work is a carefully designed and richly annotated dataset, for which we then provide an in-depth analysis by investigating characteristic distributions and correlations of the assigned labels. We reveal interesting patterns and intersections between the annotation categories and properties of our dataset, which enable insights into the characteristics of both argumentative texts and implicit knowledge in terms of structural features and semantic information. The results of our analysis can help to assist automated argument analysis and can guide the process of revealing implicit information in argumentative texts automatically.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.282}
}

@InProceedings{faralli-velardi-yusifli:2020:LREC,

```
  author    = {Faralli, Stefano  and  Velardi, Paola  and  Yusifli,
Farid},
  title     = {Multiple Knowledge GraphDB (MKGDB)},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2325--2331},
  abstract  = {We present MKGDB, a large-scale graph database
created as a combination of multiple taxonomy backbones extracted
from 5 existing knowledge graphs, namely: ConceptNet, DBpedia,
WebIsAGraph, WordNet and the Wikipedia category hierarchy. MKGDB,
thanks the versatility of the Neo4j graph database manager
technology, is intended to favour and help the development of open-
domain natural language processing applications relying on knowledge
bases, such as information extraction, hypernymy discovery, topic
clustering, and others. Our resource consists of a large hypernymy
graph which counts more than 37 million nodes and more than 81
million hypernymy relations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.283}
}

@InProceedings{morenoschneider-EtAl:2020:LREC,
  author    = {Moreno-Schneider, Julian  and  Rehm, Georg  and
Montiel-Ponsoda, Elena  and  Rodriguez-Doncel, Víctor  and  Revenko,
Artem  and  Karampatakis, Sotirios  and  Khvalchik, Maria  and
Sageder, Christian  and  Gracia, Jorge  and  Maganza, Filippo},
  title     = {Orchestrating NLP Services for the Legal Domain},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2332--2340},
  abstract  = {Legal technology is currently receiving a lot of
attention from various angles. In this contribution we describe the
main technical components of a system that is currently under
development in the European innovation project Lynx, which includes
partners from industry and research. The key contribution of this
paper is a workflow manager that enables the flexible orchestration
of workflows based on a portfolio of Natural Language Processing and
Content Curation services as well as a Multilingual Legal Knowledge
Graph that contains semantic information and meaningful references
to legal documents. We also describe different use cases with which
we experiment and develop prototypical solutions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.284}
}

@InProceedings{bordea-EtAl:2020:LREC,
  author    = {Bordea, Georgeta  and  Faralli, Stefano  and  Mougin,
Fleur  and  Buitelaar, Paul  and  Diallo, Gayo},
```

```
  title     = {Evaluation Dataset and Methodology for Extracting
Application-Specific Taxonomies from the Wikipedia Knowledge Graph},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2341--2347},
  abstract  = {In this work, we address the task of extracting
application-specific taxonomies from the category hierarchy of
Wikipedia. Previous work on pruning the Wikipedia knowledge graph
relied on silver standard taxonomies which can only be automatically
extracted for a small subset of domains rooted in relatively focused
nodes, placed at an intermediate level in the knowledge graphs. In
this work, we propose an iterative methodology to extract an
application-specific gold standard dataset from a knowledge graph
and an evaluation framework to comparatively assess the quality of
noisy automatically extracted taxonomies. We employ an existing
state of the art algorithm in an iterative manner and we propose
several sampling strategies to reduce the amount of manual work
needed for evaluation. A first gold standard dataset is released to
the research community for this task along with a companion
evaluation framework. This dataset addresses a real-world
application from the medical domain, namely the extraction of food-
drug and herb-drug interactions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.285}
}

@InProceedings{randria-EtAl:2020:LREC,
  author    = {Randria, Estelle  and  Fontan, Lionel  and  Le Coz,
Maxime  and  Ferrané, Isabelle  and  Pinquier, Julien},
  title     = {Subjective Evaluation of Comprehensibility in Movie
Interactions},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2348--2357},
  abstract  = {Various research works have dealt with the
comprehensibility of textual, audio, or audiovisual documents, and
showed that factors related to text (e.g. linguistic complexity),
sound (e.g. speech intelligibility), image (e.g. presence of visual
context), or even to cognition and emotion can play a major role in
the ability of humans to understand the semantic and pragmatic
contents of a given document. However, to date, no reference human
data is available that could help investigating the role of the
linguistic and extralinguistic information present at these
different levels (i.e., linguistic, audio/phonetic, and visual) in
multimodal documents (e.g., movies). The present work aimed at
building a corpus of human annotations that would help to study
further how much and in which way the human perception of
```

```
comprehensibility (i.e., of the difficulty of comprehension,
referred in this paper as overall difficulty) of audiovisual
documents is affected (1) by lexical complexity, grammatical
complexity, and speech intelligibility, and (2) by the modality/ies
(text, audio, video) available to the human recipient.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.286}
}

@InProceedings{lenaraz-reimerink-cabezasgarca:2020:LREC,
  author    = {León-Araúz, Pilar  and  Reimerink, Arianne  and
Cabezas-García, Melania},
  title     = {Representing Multiword Term Variation in a
Terminological Knowledge Base: a Corpus-Based Study},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2358--2367},
  abstract  = {In scientific and technical communication, multiword
terms are the most frequent type of lexical units. Rendering them in
another language is not an easy task due to their cognitive
complexity, the proliferation of different forms, and their
unsystematic representation in terminographic resources. This often
results in a broad spectrum of translations for multiword terms,
which also foment term variation since they consist of two or more
constituents. In this study we carried out a quantitative and
qualitative analysis of Spanish translation variants of a set of
environment-related concepts by evaluating equivalents in three
parallel corpora, two comparable corpora and two terminological
resources. Our results showed that MWTs exhibit a significant degree
of term variation of different characteristics, which were used to
establish a set of criteria according to which term variants should
be selected, organized and described in terminological knowledge
bases.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.287}
}

@InProceedings{dan-he-roth:2020:LREC,
  author    = {Dan, Soham  and  He, Hangfeng  and  Roth, Dan},
  title     = {Understanding Spatial Relations through Multiple
Modalities},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2368--2372},
  abstract  = {Recognizing spatial relations and reasoning about
them is essential in multiple applications including navigation,
direction giving and human-computer interaction in general. Spatial
relations between objects can either be explicit – expressed as
```

spatial prepositions, or implicit – expressed by spatial verbs such as moving, walking, shifting, etc. Both these, but implicit relations in particular, require significant common sense understanding. In this paper, we introduce the task of inferring implicit and explicit spatial relations between two entities in an image. We design a model that uses both textual and visual information to predict the spatial relations, making use of both positional and size information of objects and image embeddings. We contrast our spatial model with powerful language models and show how our modeling complements the power of these, improving prediction accuracy and coverage and facilitates dealing with unseen subjects, objects and relations.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.288}
}

@InProceedings{roy-bhatia-jain:2020:LREC,
  author   = {Roy, Dwaipayan and Bhatia, Sumit and Jain, Prateek},
  title   = {A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2373--2380},
  abstract = {Wikipedia is the largest web-based open encyclopedia covering more than three hundred languages. However, different language editions of Wikipedia differ significantly in terms of their information coverage. We present a systematic comparison of information coverage in English Wikipedia (most exhaustive) and Wikipedias in eight other widely spoken languages (Arabic, German, Hindi, Korean, Portuguese, Russian, Spanish and Turkish). We analyze the content present in the respective Wikipedias in terms of the coverage of topics as well as the depth of coverage of topics included in these Wikipedias. Our analysis quantifies and provides useful insights about the information gap that exists between different language editions of Wikipedia and offers a roadmap for the IR community to bridge this gap.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.289}
}

@InProceedings{sato-miyazawa:2020:LREC,
  author   = {Sato, Yoshinao and Miyazawa, Kouki},
  title   = {Quality Estimation for Partially Subjective Classification Tasks via Crowdsourcing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {229--235},
  abstract  = {The quality estimation of artifacts generated by
creators via crowdsourcing has great significance for the
construction of a large-scale data resource. A common approach to
this problem is to ask multiple reviewers to evaluate the same
artifacts. However, the commonly used majority voting method to
aggregate reviewers' evaluations does not work effectively for
partially subjective or purely subjective tasks because reviewers'
sensitivity and bias of evaluation tend to have a wide variety. To
overcome this difficulty, we propose a probabilistic model for
subjective classification tasks that incorporates the qualities of
artifacts as well as the abilities and biases of creators and
reviewers as latent variables to be jointly inferred. We applied
this method to the partially subjective task of speech
classification into the following four attitudes: agreement,
disagreement, stalling, and question. The result shows that the
proposed method estimates the quality of speech more effectively
than a vote aggregation, measured by correlation with a fine-grained
classification by experts.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.29}
}

@InProceedings{kmetty-EtAl:2020:LREC,
  author    = {Kmetty, Zoltán  and  Vincze, Veronika  and  Demszky,
Dorottya  and  Ring, Orsolya  and  Nagy, Balázs  and  Szabó, Martina
Katalin},
  title     = {Pártélet: A Hungarian Corpus of Propaganda Texts from
the Hungarian Socialist Era},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2381--2388},
  abstract  = {In this paper, we present Pártélet, a digitized
Hungarian corpus of Communist propaganda texts. Pártélet was the
official journal of the governing party during the Hungarian
socialism from 1956 to 1989, hence it represents the direct
political agitation and propaganda of the dictatorial system in
question. The paper has a dual purpose: first, to present a general
review of the corpus compilation process and the basic statistical
data of the corpus, and second, to demonstrate through two case
studies what the dataset can be used for. We show that our corpus
provides a unique opportunity for conducting research on Hungarian
propaganda discourse, as well as analyzing changes of this discourse
over a 35-year period of time with computer-assisted methods.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.290}
}

@InProceedings{noullet-mix-frber:2020:LREC,
  author    = {Noullet, Kristian  and  Mix, Rico  and  Färber,
Michael},
  title     = {KORE 50\^{}DYWC: An Evaluation Data Set for Entity
```

Linking Based on DBpedia, YAGO, Wikidata, and Crunchbase},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2389--2395},
  abstract = {A major domain of research in natural language processing is named entity recognition and disambiguation (NERD). One of the main ways of attempting to achieve this goal is through use of Semantic Web technologies and its structured data formats. Due to the nature of structured data, information can be extracted more easily, therewith allowing for the creation of knowledge graphs. In order to properly evaluate a NERD system, gold standard data sets are required. A plethora of different evaluation data sets exists, mostly relying on either Wikipedia or DBpedia. Therefore, we have extended a widely-used gold standard data set, KORE 50, to not only accommodate NERD tasks for DBpedia, but also for YAGO, Wikidata and Crunchbase. As such, our data set, KORE 50\^{}DYWC, allows for a broader spectrum of evaluation. Among others, the knowledge graph agnosticity of NERD systems may be evaluated which, to the best of our knowledge, was not possible until now for this number of knowledge graphs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.291}
}

@InProceedings{alacam-EtAl:2020:LREC,
  author   = {Alacam, Özge and Ruppert, Eugen and Salama, Amr Rekaby and Staron, Tobias and Menzel, Wolfgang},
  title   = {Eye4Ref: A Multimodal Eye Movement Dataset of Referentially Complex Situations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2396--2404},
  abstract = {Eye4Ref is a rich multimodal dataset of eye-movement recordings collected from referentially complex situated settings where the linguistic utterances and their visual referential world were available to the listener. It consists of not only fixation parameters but also saccadic movement parameters that are time-locked to accompanying German utterances (with English translations). Additionally, it also contains symbolic knowledge (contextual) representations of the images to map the referring expressions onto the objects in corresponding images. Overall, the data was collected from 62 participants in three different experimental setups (86 systematically controlled sentence--image pairs and 1844 eye-movement recordings). Referential complexity was controlled by visual manipulations (e.g. number of objects, visibility of the target items, etc.), and by linguistic manipulations (e.g., the position of the disambiguating word in a

sentence). This multimodal dataset, in which the three different sources of information namely eye-tracking, language, and visual environment are aligned, offers a test of various research questions not from only language perspective but also computer vision.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.292}
}

@InProceedings{chen-cao-jiang:2020:LREC,
  author   = {Chen, Jiahao and Cao, Chenjie and Jiang, Xiuyan},
  title   = {SiBert: Enhanced Chinese Pre-trained Language Model with Sentence Insertion},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2405--2412},
  abstract = {Pre-trained models have achieved great success in learning unsupervised language representations by self-supervised tasks on large-scale corpora. Recent studies mainly focus on how to fine-tune different downstream tasks from a general pre-trained model. However, some studies show that customized self-supervised tasks for a particular type of downstream task can effectively help the pre-trained model to capture more corresponding knowledge and semantic information. Hence a new pre-training task called Sentence Insertion (SI) is proposed in this paper for Chinese query-passage pairs NLP tasks including answer span prediction, retrieval question answering and sentence level cloze test. The related experiment results indicate that the proposed SI can improve the performance of the Chinese Pre-trained models significantly. Moreover, a word segmentation method called SentencePiece is utilized to further enhance Chinese Bert performance for tasks with long texts. The complete source code is available at https://github.com/ewrfcas/SiBert_tensorflow.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.293}
}

@InProceedings{roark-EtAl:2020:LREC,
  author   = {Roark, Brian and Wolf-Sonkin, Lawrence and Kirov, Christo and Mielke, Sabrina J. and Johny, Cibu and Demirsahin, Isin and Hall, Keith},
  title   = {Processing South Asian Languages Written in the Latin Script: the Dakshina Dataset},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2413--2423},
  abstract = {This paper describes the Dakshina dataset, a new resource consisting of text in both the Latin and native scripts for 12 South Asian languages. The dataset includes, for each language:

```
1) native script Wikipedia text; 2) a romanization lexicon; and 3)
full sentence parallel data in both a native script of the language
and the basic Latin alphabet. We document the methods used for
preparation and selection of the Wikipedia text in each language;
collection of attested romanizations for sampled lexicons; and
manual romanization of held-out sentences from the native script
collections. We additionally provide baseline results on several
tasks made possible by the dataset, including single word
transliteration, full sentence transliteration, and language
modeling of native script and romanized text.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.294}
}

@InProceedings{melli-EtAl:2020:LREC,
  author    = {Melli, Gabor  and  Eldallal, Abdelrhman  and  Lazem,
Bassim  and  Moreira, Olga},
  title     = {GM-RKB WikiText Error Correction Task and Baselines},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2424--2430},
  abstract  = {We introduce the GM-RKB WikiText Error Correction
Task for the automatic detection and correction of typographical
errors in WikiText annotated pages. The included corpus is based on
a snapshot of the GM-RKB domain-specific semantic wiki consisting of
a large collection of concepts, personages, and publications primary
centered on data mining and machine learning research topics.
Numerous Wikipedia pages were also included as additional training
data in the task's evaluation process. The corpus was then
automatically updated to synthetically include realistic errors to
produce a training and evaluation ground truth comparison. We
designed and evaluated two supervised baseline WikiFixer error
correction methods: (1) a naive approach based on a maximum
likelihood character-level language model; (2) and an advanced model
based on a sequence-to-sequence (seq2seq) neural network
architecture. Both error correction models operated at a character
level. When compared against an off-the-shelf word-level spell
checker these methods showed a significant improvement in the task's
performance -- with the seq2seq-based model correcting a higher
number of errors than it introduced. Finally, we published our data
and code.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.295}
}

@InProceedings{beyer-kauermann-schtze:2020:LREC,
  author    = {Beyer, Anne  and  Kauermann, Göran  and  Schütze,
Hinrich},
  title     = {Embedding Space Correlation as a Measure of Domain
Similarity},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2431--2439},
  abstract  = {Prior work has determined domain similarity using
text-based features of a corpus. However, when using pre-trained
word embeddings, the underlying text corpus might not be accessible
anymore. Therefore, we propose the CCA measure, a new measure of
domain similarity based directly on the dimension-wise correlations
between corresponding embedding spaces. Our results suggest that an
inherent notion of domain can be captured this way, as we are able
to reproduce our findings for different domain comparisons for
English, German, Spanish and Czech as well as in cross-lingual
comparisons. We further find a threshold at which the CCA measure
indicates that two corpora come from the same domain in a
monolingual setting by applying permutation tests. By evaluating the
usability of the CCA measure in a domain adaptation application, we
also show that it can be used to determine which corpora are more
similar to each other in a cross-domain sentiment detection task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.296}
}

@InProceedings{guo-EtAl:2020:LREC,
  author    = {Guo, Mandy  and  Dai, Zihang  and  Vrandečić, Denny
and  Al-Rfou, Rami},
  title     = {Wiki-40B: Multilingual Language Model Dataset},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2440--2452},
  abstract  = {We propose a new multilingual language model
benchmark that is composed of 40+ languages spanning several scripts
and linguistic families. With around 40 billion characters, we hope
this new resource will accelerate the research of multilingual
modeling. We train monolingual causal language models using a state-
of-the-art model (Transformer-XL) establishing baselines for many
languages. We also introduce the task of multilingual causal
language modeling where we train our model on the combined text of
40+ languages from Wikipedia with different vocabulary sizes and
evaluate on the languages individually. We released the cleaned-up
text of 40+ Wikipedia language editions, the corresponding trained
monolingual language models, and several multilingual language
models with different fixed vocabulary sizes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.297}
}

@InProceedings{sharoff:2020:LREC,
  author    = {Sharoff, Serge},
  title     = {Know thy Corpus! Robust Methods for Digital Curation
of Web corpora},
```

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2453--2460},
  abstract  = {This paper proposes a novel framework for digital
curation of Web corpora in order to provide robust estimation of
their parameters, such as their composition and the lexicon. In
recent years language models pre-trained on large corpora emerged as
clear winners in numerous NLP tasks, but no proper analysis of the
corpora which led to their success has been conducted. The paper
presents a procedure for robust frequency estimation, which helps in
establishing the core lexicon for a given corpus, as well as a
procedure for estimating the corpus composition via unsupervised
topic models and via supervised genre classification of Web pages.
The results of the digital curation study applied to several Web-
derived corpora demonstrate their considerable differences. First,
this concerns different frequency bursts which impact the core
lexicon obtained from each corpus. Second, this concerns the kinds
of texts they contain. For example, OpenWebText contains
considerably more topical news and political argumentation in
comparison to ukWac or Wikipedia. The tools and the results of
analysis have been released.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.298}
}

@InProceedings{king-cook:2020:LREC,
  author    = {King, Milton  and  Cook, Paul},
  title     = {Evaluating Approaches to Personalizing Language
Models},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2461--2469},
  abstract  = {In this work, we consider the problem of
personalizing language models, that is, building language models
that are tailored to the writing style of an individual. Because
training language models requires a large amount of text, and
individuals do not necessarily possess a large corpus of their
writing that could be used for training, approaches to personalizing
language models must be able to rely on only a small amount of text
from any one user. In this work, we compare three approaches to
personalizing a language model that was trained on a large
background corpus using a relatively small amount of text from an
individual user. We evaluate these approaches using perplexity, as
well as two measures based on next word prediction for smartphone
soft keyboards. Our results show that when only a small amount of
user-specific text is available, an approach based on priming gives
the most improvement, while when larger amounts of user-specific
```

```
text are available, an approach based on language model
interpolation performs best. We carry out further experiments to
show that these approaches to personalization outperform language
model adaptation based on demographic factors.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.299}
}

@InProceedings{bernard-han:2020:LREC,
  author    = {Bernard, Timothée  and  Han, Ting},
  title     = {Mandarinograd: A Chinese Collection of Winograd
Schemas},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {21--26},
  abstract  = {This article introduces Mandarinograd, a corpus of
Winograd Schemas in Mandarin Chinese. Winograd Schemas are
particularly challenging anaphora resolution problems, designed to
involve common sense reasoning and to limit the biases and artefacts
commonly found in natural language understanding datasets.
Mandarinograd contains the schemas in their traditional form, but
also as natural language inference instances (ENTAILMENT or NO
ENTAILMENT pairs) as well as in their fully disambiguated candidate
forms. These two alternative representations are often used by
modern solvers but existing datasets present automatically converted
items that sometimes contain syntactic or semantic anomalies. We
detail the difficulties faced when building this corpus and explain
how weavoided the anomalies just mentioned. We also show that
Mandarinograd is resistant to a statistical method based on a
measure of word association.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.3}
}

@InProceedings{hahm-EtAl:2020:LREC,
  author    = {Hahm, Younggyun  and  Noh, Youngbin  and  Han, Ji
Yoon  and  Oh, Tae Hwan  and  Choe, Hyonsu  and  Kim, Hansaem  and
Choi, Key-Sun},
  title     = {Crowdsourcing in the Development of a Multilingual
FrameNet: A Case Study of Korean FrameNet},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {236--244},
  abstract  = {Using current methods, the construction of
multilingual resources in FrameNet is an expensive and complex task.
While crowdsourcing is a viable alternative, it is difficult to
include non-native English speakers in such efforts as they often
have difficulty with English-based FrameNet tools. In this work, we
```

investigated cross-lingual issues in crowdsourcing approaches for multilingual FrameNets, specifically in the context of the newly constructed Korean FrameNet. To accomplish this, we evaluated the effectiveness of various crowdsourcing settings whereby certain types of information are provided to workers, such as English definitions in FrameNet or translated definitions. We then evaluated whether the crowdsourced results accurately captured the meaning of frames both cross-culturally and cross-linguistically, and found that by allowing the crowd workers to make intuitive choices, they achieved a quality comparable to that of trained FrameNet experts (F1 > 0.75). The outcomes of this work are now publicly available as a new release of Korean FrameNet 1.1.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.30}
}

@InProceedings{kipyatkova-karpov:2020:LREC,
  author   = {Kipyatkova, Irina and Karpov, Alexey},
  title   = {Class-based LSTM Russian Language Model with Linguistic Information},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2470--2474},
  abstract = {In the paper, we present class-based LSTM Russian language models (LMs) with classes generated with the use of both word frequency and linguistic information data, obtained with the help of the “VisualSynan” software from the AOT project. We have created LSTM LMs with various numbers of classes and compared them with word-based LM and class-based LM with word2vec class generation in terms of perplexity, training time, and WER. In addition, we performed a linear interpolation of LSTM language models with the baseline 3-gram language model. The LSTM language models were used for very large vocabulary continuous Russian speech recognition at an N-best list rescoring stage. We achieved significant progress in training time reduction with only slight degradation in recognition accuracy comparing to the word-based LM. In addition, our LM with classes generated using linguistic information outperformed LM with classes generated using word2vec. We achieved WER of 14.94 \% at our own speech corpus of continuous Russian speech that is 15 \% relative reduction with respect to the baseline 3-gram model.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.300}
}

@InProceedings{ralethe:2020:LREC,
  author   = {Ralethe, Sello},
  title   = {Adaptation of Deep Bidirectional Transformers for Afrikaans Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2475--2478},
  abstract  = {The recent success of pretrained language models in
Natural Language Processing has sparked interest in training such
models for languages other than English. Currently, training of
these models can either be monolingual or multilingual based. In the
case of multilingual models, such models are trained on concatenated
data of multiple languages. We introduce AfriBERT, a language model
for the Afrikaans language based on Bidirectional Encoder
Representation from Transformers (BERT). We compare the performance
of AfriBERT against multilingual BERT in multiple downstream tasks,
namely part-of-speech tagging, named-entity recognition, and
dependency parsing. Our results show that AfriBERT improves the
current state-of-the-art in most of the tasks we considered, and
that transfer learning from multilingual to monolingual model can
have a significant performance improvement on downstream tasks. We
release the pretrained model for AfriBERT.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.301}
}

@InProceedings{le-EtAl:2020:LREC,
  author    = {Le, Hang  and  Vial, Loïc  and  Frej, Jibril  and
Segonne, Vincent  and  Coavoux, Maximin  and  Lecouteux, Benjamin
and  Allauzen, Alexandre  and  Crabbé, Benoit  and  Besacier,
Laurent  and  Schwab, Didier},
  title     = {FlauBERT: Unsupervised Language Model Pre-training
for French},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2479--2490},
  abstract  = {Language models have become a key step to achieve
state-of-the art results in many different Natural Language
Processing (NLP) tasks. Leveraging the huge amount of unlabeled
texts nowadays available, they provide an efficient way to pre-train
continuous word representations that can be fine-tuned for a
downstream task, along with their contextualization at the sentence
level. This has been widely demonstrated for English using
contextualized representations (Dai and Le, 2015; Peters et al.,
2018; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al.,
2019; Yang et al., 2019b). In this paper, we introduce and share
FlauBERT, a model learned on a very large and heterogeneous French
corpus. Models of different sizes are trained using the new CNRS
(French National Centre for Scientific Research) Jean Zay
supercomputer. We apply our French language models to diverse NLP
tasks (text classification, paraphrasing, natural language
inference, parsing, word sense disambiguation) and show that most of
the time they outperform other pre-training approaches. Different
versions of FlauBERT as well as a unified evaluation protocol for
the downstream tasks, called FLUE (French Language Understanding
```

Evaluation), are shared to the research community for further reproducible experiments in French NLP.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.302}
}

@InProceedings{ciosici-assent-derczynski:2020:LREC,
  author   = {Ciosici, Manuel R. and Assent, Ira and Derczynski, Leon},
  title   = {Accelerated High-Quality Mutual-Information Based Word Clustering},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2491--2496},
  abstract = {Word clustering groups words that exhibit similar properties. One popular method for this is Brown clustering, which uses short-range distributional information to construct clusters. Specifically, this is a hard hierarchical clustering with a fixed-width beam that employs bi-grams and greedily minimizes global mutual information loss. The result is word clusters that tend to outperform or complement other word representations, especially when constrained by small datasets. However, Brown clustering has high computational complexity and does not lend itself to parallel computation. This, together with the lack of efficient implementations, limits their applicability in NLP. We present efficient implementations of Brown clustering and the alternative Exchange clustering as well as a number of methods to accelerate the computation of both hierarchical and flat clusters. We show empirically that clusters obtained with the accelerated method match the performance of clusters computed using the original methods.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.303}
}

@InProceedings{coulange-rossato:2020:LREC,
  author   = {Coulange, Sylvain and Rossato, Solange},
  title   = {Rhythmic Proximity Between Natives And Learners Of French - Evaluation of a metric based on the CEFC corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2497--2502},
  abstract = {This work aims to better understand the role of rhythm in foreign accent, and its modelling. We made a model of rhythm in French taking into account its variability, thanks to the Corpus pour l'Étude du Français Contemporain (CEFC), which contains up to 300 hours of speech of a wide variety of speaker profiles and situations. 16 parameters were computed, each of them being based on segment duration, such as voicing and intersyllabic timing. All the

parameters are fully automatically detected from signal, without ASR or transcription. A gaussian mixture model was trained on 1,340 native speakers of French; any 30-second minimum speech may be computed to get the probability of its belonging to this model. We tested it with 146 test native speakers (NS), 37 non-native speakers (NNS) from the same corpus, and 29 non-native Japanese learners of French (JpNNS) from an independent corpus. The probability of NNS having inferior log-likelihood to NS was only a tendency (p=.067), maybe due to the heterogeneity of French proficiency of the speakers; but a much bigger probability was obtained for JpNNS (p<. 0001), where all speakers were A2 level. Eta-squared test showed that most efficient parameters were intersyllabic mean duration and variation coefficient, along with speech rate for NNS; and speech rate and phonation ratio for JpNNS.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.304}
}

@InProceedings{speranza-EtAl:2020:LREC,
  author   = {Speranza, Giulia and di Buono, Maria Pia and Monti, Johanna and Sangati, Federico},
  title   = {From Linguistic Resources to Ontology-Aware Terminologies: Minding the Representation Gap},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2503--2510},
  abstract = {Terminological resources have proven crucial in many applications ranging from Computer-Aided Translation tools to authoring softwares and multilingual and cross-lingual information retrieval systems. Nonetheless, with the exception of a few felicitous examples, such as the IATE (Interactive Terminology for Europe) Termbank, many terminological resources are not available in standard formats, such as Term Base eXchange (TBX), thus preventing their sharing and reuse. Yet, these terminologies could be improved associating the correspondent ontology-based information. The research described in the present contribution demonstrates the process and the methodologies adopted in the automatic conversion into TBX of such type of resources, together with their semantic enrichment based on the formalization of ontological information into terminologies. We present a proof-of-concept using the Italian Linguistic Resource for the Archaeological domain (developed according to Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation). Further, we introduce the conversion tool developed to support the process of creating ontology-aware terminologies for improving interoperability and sharing of existing language technologies and data sets.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.305}
}

@InProceedings{arslan-EtAl:2020:LREC,
  author   = {Arslan, Fatma and Caraballo, Josue and Jimenez,

```
Damian  and  Li, Chengkai},
  title     = {Modeling Factual Claims with Semantic Frames},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2511--2520},
  abstract  = {In this paper, we introduce an extension of the
Berkeley FrameNet for the structured and semantic modeling of
factual claims. Modeling is a robust tool that can be leveraged in
many different tasks such as matching claims to existing fact-checks
and translating claims to structured queries. Our work introduces 11
new manually crafted frames along with 9 existing FrameNet frames,
all of which have been selected with fact-checking in mind. Along
with these frames, we are also providing 2,540 fully annotated
sentences, which can be used to understand how these frames are
intended to work and to train machine learning models. Finally, we
are also releasing our annotation tool to facilitate other
researchers to make their own local extensions to FrameNet.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.306}
}

@InProceedings{gupta-boulianne:2020:LREC,
  author    = {Gupta, Vishwa  and  Boulianne, Gilles},
  title     = {Automatic Transcription Challenges for Inuktitut, a
Low-Resource Polysynthetic Language},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2521--2527},
  abstract  = {We introduce the first attempt at automatic speech
recognition (ASR) in Inuktitut, as a representative for
polysynthetic, low-resource languages, like many of the 900
Indigenous languages spoken in the Americas. As most previous work
on Inuktitut, we use texts from parliament proceedings, but in
addition we have access to 23 hours of transcribed oral stories.
With this corpus, we show that Inuktitut displays a much higher
degree of polysynthesis than other agglutinative languages usually
considered in ASR, such as Finnish or Turkish. Even with a
vocabulary of 1.3 million words derived from proceedings and
stories, held-out stories have more than 60\% of words out-of-
vocabulary. We train bi-directional LSTM acoustic models, then
investigate word and subword units, morphemes and syllables, and a
deep neural network that finds word boundaries in subword sequences.
We show that acoustic decoding using syllables decorated with word
boundary markers results in the lowest word error rate.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.307}
}
```

```
@InProceedings{dunn-adams:2020:LREC,
  author    = {Dunn, Jonathan  and  Adams, Ben},
  title     = {Geographically-Balanced Gigaword Corpora for 50
Language Varieties},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2528--2536},
  abstract  = {While text corpora have been steadily increasing in
overall size, even very large corpora are not designed to represent
global population demographics. For example, recent work has shown
that existing English gigaword corpora over-represent inner-circle
varieties from the US and the UK. To correct implicit geographic and
demographic biases, this paper uses country-level population
demographics to guide the construction of gigaword web corpora. The
resulting corpora explicitly match the ground-truth geographic
distribution of each language, thus equally representing language
users from around the world. This is important because it ensures
that speakers of under-resourced language varieties (i.e., Indian
English or Algerian French) are represented, both in the corpora
themselves but also in derivative resources like word embeddings.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.308}
}

@InProceedings{amjad-sidorov-zhila:2020:LREC,
  author    = {Amjad, Maaz  and  Sidorov, Grigori  and  Zhila,
Alisa},
  title     = {Data Augmentation using Machine Translation for Fake
News Detection in the Urdu Language},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2537--2542},
  abstract  = {The task of fake news detection is to distinguish
legitimate news articles that describe real facts from those which
convey deceiving and fictitious information. As the fake news
phenomenon is omnipresent across all languages, it is crucial to be
able to efficiently solve this problem for languages other than
English. A common approach to this task is supervised classification
using features of various complexity. Yet supervised machine
learning requires substantial amount of annotated data. For English
and a small number of other languages, annotated data availability
is much higher, whereas for the vast majority of languages, it is
almost scarce. We investigate whether machine translation at its
present state could be successfully used as an automated technique
for annotated corpora creation and augmentation for fake news
detection focusing on the English-Urdu language pair. We train a
fake news classifier for Urdu on (1) the manually annotated dataset
```

originally in Urdu and (2) the machine-translated version of an existing annotated fake news dataset originally in English. We show that at the present state of machine translation quality for the English-Urdu language pair, the fully automated data augmentation through machine translation did not provide improvement for fake news detection in Urdu.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.309}
}

@InProceedings{iskender-polzehl-mller:2020:LREC,
  author   = {Iskender, Neslihan and Polzehl, Tim and Möller, Sebastian},
  title   = {Towards a Reliable and Robust Methodology for Crowd-Based Subjective Quality Assessment of Query-Based Extractive Text Summarization},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {245--253},
  abstract = {The intrinsic and extrinsic quality evaluation is an essential part of the summary evaluation methodology usually conducted in a traditional controlled laboratory environment. However, processing large text corpora using these methods reveals expensive from both the organizational and the financial perspective. For the first time, and as a fast, scalable, and cost-effective alternative, we propose micro-task crowdsourcing to evaluate both the intrinsic and extrinsic quality of query-based extractive text summaries. To investigate the appropriateness of crowdsourcing for this task, we conduct intensive comparative crowdsourcing and laboratory experiments, evaluating nine extrinsic and intrinsic quality measures on 5-point MOS scales. Correlating results of crowd and laboratory ratings reveals high applicability of crowdsourcing for the factors overall quality, grammaticality, non-redundancy, referential clarity, focus, structure \& coherence, summary usefulness, and summary informativeness. Further, we investigate the effect of the number of repetitions of assessments on the robustness of mean opinion score of crowd ratings, measured against the increase of correlation coefficients between crowd and laboratory. Our results suggest that the optimal number of repetitions in crowdsourcing setups, in which any additional repetitions do no longer cause an adequate increase of overall correlation coefficients, lies between seven and nine for intrinsic and extrinsic quality factors.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.31}
}

@InProceedings{outsios-EtAl:2020:LREC,
  author   = {Outsios, Stamatis and Karatsalos, Christos and Skianis, Konstantinos and Vazirgiannis, Michalis},
  title   = {Evaluation of Greek Word Embeddings},
  booktitle   = {Proceedings of The 12th Language Resources and

```
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2543--2551},
  abstract  = {Since word embeddings have been the most popular
input for many NLP tasks, evaluating their quality is critical. Most
research efforts are focusing on English word embeddings. This paper
addresses the problem of training and evaluating such models for the
Greek language. We present a new word analogy test set considering
the original English Word2vec analogy test set and some specific
linguistic aspects of the Greek language as well. Moreover, we
create a Greek version of WordSim353 test collection for a basic
evaluation of word similarities. Produced resources are available
for download. We test seven word vector models and our evaluation
shows that we are able to create meaningful representations. Last,
we discover that the morphological complexity of the Greek language
and polysemy can influence the quality of the resulting word
embeddings.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.310}
}

@InProceedings{papavassiliou-owens-kosmopoulos:2020:LREC,
  author    = {Papavassiliou, Katerina  and  Owens, Gareth  and
Kosmopoulos, Dimitrios},
  title     = {A Dataset of Mycenaean Linear B Sequences},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2552--2561},
  abstract  = {We present our work towards a dataset of Mycenaean
Linear B sequences gathered from the Mycenaean inscriptions written
in the 13th and 14th century B.C. (c. 1400-1200 B.C.). The dataset
contains sequences of Mycenaean words and ideograms according to the
rules of the Mycenaean Greek language in the Late Bronze Age. Our
ultimate goal is to contribute to the study, reading and
understanding of ancient scripts and languages. Focusing on
sequences, we seek to exploit the structure of the entire language,
not just the Mycenaean vocabulary, to analyse sequential patterns.
We use the dataset to experiment on estimating the missing symbols
in damaged inscriptions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.311}
}

@InProceedings{joanis-EtAl:2020:LREC,
  author    = {Joanis, Eric  and  Knowles, Rebecca  and  Kuhn,
Roland  and  Larkin, Samuel  and  Littell, Patrick  and  Lo, Chi-kiu
and  Stewart, Darlene  and  Micher, Jeffrey},
  title     = {The Nunavut Hansard Inuktitut–English Parallel Corpus
3.0 with Preliminary Machine Translation Results},
```

  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2562--2572},
  abstract = {The Inuktitut language, a member of the Inuit-Yupik-Unangan language family, is spoken across Arctic Canada and noted for its morphological complexity. It is an official language of two territories, Nunavut and the Northwest Territories, and has recognition in additional regions. This paper describes a newly released sentence-aligned Inuktitut–English corpus based on the proceedings of the Legislative Assembly of Nunavut, covering sessions from April 1999 to June 2017. With approximately 1.3 million aligned sentence pairs, this is, to our knowledge, the largest parallel corpus of a polysynthetic language or an Indigenous language of the Americas released to date. The paper describes the alignment methodology used, the evaluation of the alignments, and preliminary experiments on statistical and neural machine translation (SMT and NMT) between Inuktitut and English, in both directions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.312}
}

@InProceedings{michel-hangya-fraser:2020:LREC,
  author   = {Michel, Leah and Hangya, Viktor and Fraser, Alexander},
  title   = {Exploring Bilingual Word Embeddings for Hiligaynon, a Low-Resource Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2573--2580},
  abstract = {This paper investigates the use of bilingual word embeddings for mining Hiligaynon translations of English words. There is very little research on Hiligaynon, an extremely low-resource language of Malayo-Polynesian origin with over 9 million speakers in the Philippines (we found just one paper). We use a publicly available Hiligaynon corpus with only 300K words, and match it with a comparable corpus in English. As there are no bilingual resources available, we manually develop a English-Hiligaynon lexicon and use this to train bilingual word embeddings. But we fail to mine accurate translations due to the small amount of data. To find out if the same holds true for a related language pair, we simulate the same low-resource setup on English to German and arrive at similar results. We then vary the size of the comparable English and German corpora to determine the minimum corpus size necessary to achieve competitive results. Further, we investigate the role of the seed lexicon. We show that with the same corpus size but with a smaller seed lexicon, performance can surpass results of previous

studies. We release the lexicon of 1,200 English-Hiligaynon word pairs we created to encourage further investigation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.313}
}

@InProceedings{zueva-kuznetsova-tyers:2020:LREC,
  author   = {Zueva, Anna and Kuznetsova, Anastasia and Tyers, Francis},
  title   = {A Finite-State Morphological Analyser for Evenki},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2581--2589},
  abstract = {It has been widely admitted that morphological analysis is an important step in automated text processing for morphologically rich languages. Evenki is a language with rich morphology, therefore a morphological analyser is highly desirable for processing Evenki texts and developing applications for Evenki. Although two morphological analysers for Evenki have already been developed, they are able to analyse less than a half of the available Evenki corpora. The aim of this paper is to create a new morphological analyser for Evenki. It is implemented using the Helsinki Finite-State Transducer toolkit (HFST). The lexc formalism is used to specify the morphotactic rules, which define the valid orderings of morphemes in a word. Morphophonological alternations and orthographic rules are described using the twol formalism. The lexicon is extracted from available machine-readable dictionaries. Since a part of the corpora belongs to texts in Evenki dialects, a version of the analyser with relaxed rules is developed for processing dialectal features. We evaluate the analyser on available Evenki corpora and estimate precision, recall and F-score. We obtain coverage scores of between 61\% and 87\% on the available Evenki corpora.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.314}
}

@InProceedings{mersha-wu:2020:LREC,
  author   = {Mersha, Amanuel and Wu, Stephen},
  title   = {Morphology-rich Alphasyllabary Embeddings},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2590--2595},
  abstract = {Word embeddings have been successfully trained in many languages. However, both intrinsic and extrinsic metrics are variable across languages, especially for languages that depart significantly from English in morphology and orthography. This study focuses on building a word embedding model suitable for the Semitic

```
language of Amharic (Ethiopia), which is both morphologically rich
and written as an alphasyllabary (abugida) rather than an alphabet.
We compare embeddings from tailored neural models, simple pre-
processing steps, off-the-shelf baselines, and parallel tasks on a
better-resourced Semitic language – Arabic. Experiments show our
model's performance on word analogy tasks, illustrating the
divergent objectives of morphological vs. semantic analogies.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.315}
}

@InProceedings{cruz-tan-cheng:2020:LREC,
  author    = {Cruz, Jan Christian Blaise  and  Tan, Julianne Agatha
and  Cheng, Charibeth},
  title     = {Localization of Fake News Detection via Multitask
Transfer Learning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2596--2604},
  abstract  = {The use of the internet as a fast medium of spreading
fake news reinforces the need for computational tools that combat
it. Techniques that train fake news classifiers exist, but they all
assume an abundance of resources including large labeled datasets
and expert-curated corpora, which low-resource languages may not
have. In this work, we make two main contributions: First, we
alleviate resource scarcity by constructing the first expertly-
curated benchmark dataset for fake news detection in Filipino, which
we call "Fake News Filipino." Second, we benchmark Transfer Learning
(TL) techniques and show that they can be used to train robust fake
news classifiers from little data, achieving 91\% accuracy on our
fake news dataset, reducing the error by 14\% compared to
established few-shot baselines. Furthermore, lifting ideas from
multitask learning, we show that augmenting transformer-based
transfer techniques with auxiliary language modeling losses improves
their performance by adapting to writing style. Using this, we
improve TL performance by 4-6\%, achieving an accuracy of 96\% on
our best model. Lastly, we show that our method generalizes well to
different types of news articles, including political news,
entertainment news, and opinion articles.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.316}
}

@InProceedings{casanova-EtAl:2020:LREC,
  author    = {Casanova, Edresson  and  Treviso, Marcos  and
Hübner, Lilian  and  Aluísio, Sandra},
  title     = {Evaluating Sentence Segmentation in Different
Datasets of Neuropsychological Language Tests in Brazilian
Portuguese},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {2605--2614},
  abstract  = {Automatic analysis of connected speech by natural
language processing techniques is a promising direction for
diagnosing cognitive impairments. However, some difficulties still
remain: the time required for manual narrative transcription and the
decision on how transcripts should be divided into sentences for
successful application of parsers used in metrics, such as Idea
Density, to analyze the transcripts. The main goal of this paper was
to develop a generic segmentation system for narratives of
neuropsychological language tests. We explored the performance of
our previous single-dataset-trained sentence segmentation
architecture in a richer scenario involving three new datasets used
to diagnose cognitive impairments, comprising different stories and
two types of stimulus presentation for eliciting narratives ---
visual and oral --- via illustrated story-book and sequence of
scenes, and by retelling. Also, we proposed and evaluated three
modifications to our previous RCNN architecture: (i) the inclusion
of a Linear Chain CRF; (ii) the inclusion of a self-attention
mechanism; and (iii) the replacement of the LSTM recurrent layer by
a Quasi-Recurrent Neural Network layer. Our study allowed us to
develop two new models for segmenting impaired speech
transcriptions, along with an ideal combination of datasets and
specific groups of narratives to be used as the training set.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.317}
}

@InProceedings{park-choe-ham:2020:LREC,
  author    = {Park, Kyubyong  and  Choe, Yo Joong  and  Ham,
Jiyeon},
  title     = {Jejueo Datasets for Machine Translation and Speech
Synthesis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2615--2621},
  abstract  = {Jejueo was classified as critically endangered by
UNESCO in 2010. Although diverse efforts to revitalize it have been
made, there have been few computational approaches. Motivated by
this, we construct two new Jejueo datasets: Jejueo Interview
Transcripts (JIT) and Jejueo Single Speaker Speech (JSS). The JIT
dataset is a parallel corpus containing 170k+ Jejueo-Korean
sentences, and the JSS dataset consists of 10k high-quality audio
files recorded by a native Jejueo speaker and a transcript file.
Subsequently, we build neural systems of machine translation and
speech synthesis using them. All resources are publicly available
via our GitHub repository. We hope that these datasets will attract
interest of both language and machine learning communities.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.318}
```

```
}

@InProceedings{matsuura-EtAl:2020:LREC,
  author    = {Matsuura, Kohei  and  Ueno, Sei  and  Mimura, Masato
and  Sakai, Shinsuke  and  Kawahara, Tatsuya},
  title     = {Speech Corpus of Ainu Folklore and End-to-end Speech
Recognition for Ainu Language},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2622--2628},
  abstract  = {Ainu is an unwritten language that has been spoken by
Ainu people who are one of the ethnic groups in Japan. It is
recognized as critically endangered by UNESCO and archiving and
documentation of its language heritage is of paramount importance.
Although a considerable amount of voice recordings of Ainu folklore
has been produced and accumulated to save their culture, only a
quite limited parts of them are transcribed so far. Thus, we started
a project of automatic speech recognition (ASR) for the Ainu
language in order to contribute to the development of annotated
language archives. In this paper, we report speech corpus
development and the structure and performance of end-to-end ASR for
Ainu. We investigated four modeling units (phone, syllable, word
piece, and word) and found that the syllable-based model performed
best in terms of both word and phone recognition accuracy, which
were about 60\% and over 85\% respectively in speaker-open
condition. Furthermore, word and phone accuracy of 80\% and 90\% has
been achieved in a speaker-closed setting. We also found out that a
multilingual ASR training with additional speech corpora of English
and Japanese further improves the speaker-open test accuracy.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.319}
}

@InProceedings{radhakrishnan:2020:LREC,
  author    = {Radhakrishnan, Priya},
  title     = {A Seed Corpus of Hindu Temples in India},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {254--258},
  abstract  = {Temples are an integral part of culture and heritage
of India and are centers of religious practice for practicing
Hindus. A scientific study of temples can reveal valuable insights
into Indian culture and heritage. However to the best of our
knowledge, learning resources that aid such a study are either not
publicly available or non-existent. In this endeavour we present our
initial efforts to create a corpus of Hindu temples in India. In
this paper, we present a simple, re-usable platform that creates
```

temple corpus from web text on temples. Curation is improved using classifiers trained on textual data in Wikipedia articles on Hindu temples. The training data is verified by human volunteers. The temple corpus consists of 4933 high accuracy facts about 573 temples. We make the corpus and the platform freely available. We also test the re-usability of the platform by creating a corpus of museums in India. We believe the temple corpus will aid scientific study of temples and the platform will aid in construction of similar corpuses. We believe both these will significantly contribute in promoting research on culture and heritage of a region.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.32}
}

@InProceedings{chiruzzo-EtAl:2020:LREC,
  author   = {Chiruzzo, Luis and Amarilla, Pedro and Ríos, Adolfo and Giménez Lugo, Gustavo},
  title   = {Development of a Guarani - Spanish Parallel Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2629--2633},
  abstract = {This paper presents the development of a Guarani - Spanish parallel corpus with sentence-level alignment. The Guarani sentences of the corpus use the Jopara Guarani dialect, the dialect of Guarani spoken in Paraguay, which is based on Guarani grammar and may include several Spanish loanwords or neologisms. The corpus has around 14,500 sentence pairs aligned using a semi-automatic process, containing 228,000 Guarani tokens and 336,000 Spanish tokens extracted from web sources.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.320}
}

@InProceedings{ouahrani-bennouar:2020:LREC,
  author   = {Ouahrani, Leila and Bennouar, Djamal},
  title   = {AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2634--2643},
  abstract = {Automatic short answer grading is a significant problem in E-assessment. Several models have been proposed to deal with it. Evaluation and comparison of such solutions need the availability of Datasets with manual examples. In this paper, we introduce AR-ASAG, an Arabic Dataset for automatic short answer grading. The Dataset contains 2133 pairs of (Model Answer, Student Answer) in several versions (txt, xml, Moodle xml and .db). We

```
explore then an unsupervised corpus based approach for automatic
grading adapted to the Arabic Language. We use COALS (Correlated
Occurrence Analogue to Lexical Semantic) algorithm to create
semantic space for word distribution. The summation vector model is
combined to term weighting and common words to achieve similarity
between a teacher model answer and a student answer. The approach is
particularly suitable for languages with scarce resources such as
Arabic language where robust specific resources are not yet
available. A set of experiments were conducted to analyze the effect
of domain specificity, semantic space dimension and stemming
techniques on the effectiveness of the grading model. The proposed
approach gives promising results for Arabic language. The reported
results may serve as baseline for future research work evaluation},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.321}
}

@InProceedings{ferger:2020:LREC,
  author    = {Ferger, Anne},
  title     = {Processing Language Resources of Under-Resourced and
Endangered Languages for the Generation of Augmentative Alternative
Communication Boards},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2644--2648},
  abstract  = {Under-resourced and endangered or small languages
yield problems for automatic processing and exploiting because of
the small amount of available data. This paper shows an approach
using different annotations of enriched linguistic research data to
create communication boards commonly used in Alternative
Augmentative Communication (AAC). Using manually created lexical
analysis and rich annotation (instead of high data quantity) allows
for an automated creation of AAC communication boards. The example
presented in this paper uses data of the indigenous language Dolgan
(an endangered Turkic language of Northern Siberia) created in the
project INEL(Arkhipov and Däbritz, 2018) to generate a basic
communication board with audio snippets to be used in e.g. hospital
communication or for multilingual settings. The created boards can
be importet into various AAC software. In addition, the usage of
standard formats makes this approach applicable to various different
use cases.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.322}
}

@InProceedings{aznar-gala:2020:LREC,
  author    = {Aznar, Jocelyn  and  Gala, Núria},
  title     = {The Nisvai Corpus of Oral Narrative Practices from
Malekula (Vanuatu) and its Associated Language Resources},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {2649--2656},
  abstract  = {In this paper, we present a corpus of oral narratives
from the Nisvai linguistic community and four associated language
resources. Nisvai is an oral language spoken by 200 native speakers
in the South-East of Malekula, an island of Vanuatu, Oceania. This
language had never been the focus of a research before the one
leading to this article. The corpus we present is made of 32
annotated narratives segmented into intonation units. The audio
records were transcribed using the written conventions specifically
developed for the language and translated into French. Four
associated language resources have been generated by organizing the
annotations into written documents: two of them are available online
and two in paper format. The online resources allow the users to
listen to the audio recordings whilereading the annotations. They
were built to share the results of our fieldwork and to communicate
on the Nisvai narrative practices with the researchers as well as
with a more general audience. The bilingual paper resources, a
booklet of narratives and a Nisvai-French French-Nisvai lexicon,
were designed for the Nisvai community by taking into account their
future uses (i.e. primary school).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.323}
}

@InProceedings{paschen-EtAl:2020:LREC,
  author    = {Paschen, Ludger  and  Delafontaine, François  and
Draxler, Christoph  and  Fuchs, Susanne  and  Stave, Matthew  and
Seifart, Frank},
  title     = {Building a Time-Aligned Cross-Linguistic Reference
Corpus from Language Documentation Data (DoReCo)},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2657--2666},
  abstract  = {Natural speech data on many languages have been
collected by language documentation projects aiming to preserve
lingustic and cultural traditions in audivisual records. These data
hold great potential for large-scale cross-linguistic research into
phonetics and language processing. Major obstacles to utilizing such
data for typological studies include the non-homogenous nature of
file formats and annotation conventions found both across and within
archived collections. Moreover, time-aligned audio transcriptions
are typically only available at the level of broad (multi-word)
phrases but not at the word and segment levels. We report on
solutions developed for these issues within the DoReCo
(DOcumentation REference COrpus) project. DoReCo aims at providing
time-aligned transcriptions for at least 50 collections of under-
resourced languages. This paper gives a preliminary overview of the
current state of the project and details our workflow, in particular
```

```
standardization of formats and conventions, the addition of
segmental alignments with WebMAUS, and DoReCo's applicability for
subsequent research programs. By making the data accessible to the
scientific community, DoReCo is designed to bridge the gap between
language documentation and linguistic inquiry.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.324}
}

@InProceedings{duh-EtAl:2020:LREC,
  author    = {Duh, Kevin  and  McNamee, Paul  and  Post, Matt  and
Thompson, Brian},
  title     = {Benchmarking Neural and Statistical Machine
Translation on Low-Resource African Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2667--2675},
  abstract  = {Research in machine translation (MT) is developing at
a rapid pace. However, most work in the community has focused on
languages where large amounts of digital resources are available. In
this study, we benchmark state of the art statistical and neural
machine translation systems on two African languages which do not
have large amounts of resources: Somali and Swahili. These languages
are of social importance and serve as test-beds for developing
technologies that perform reasonably well despite the low-resource
constraint. Our findings suggest that statistical machine
translation (SMT) and neural machine translation (NMT) can perform
similarly in low-resource scenarios, but neural systems require more
careful tuning to match performance. We also investigate how to
exploit additional data, such as bilingual text harvested from the
web, or user dictionaries; we find that NMT can significantly
improve in performance with the use of these additional data.
Finally, we survey the landscape of machine translation resources
for the languages of Africa and provide some suggestions for
promising future research directions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.325}
}

@InProceedings{chen-park-schwartz:2020:LREC,
  author    = {Chen, Emily  and  Park, Hyunji Hayley  and  Schwartz,
Lane},
  title     = {Improved Finite-State Morphological Analysis for St.
Lawrence Island Yupik Using Paradigm Function Morphology},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2676--2684},
  abstract  = {St. Lawrence Island Yupik is an endangered
```

polysynthetic language of the Bering Strait region. While conducting linguistic fieldwork between 2016 and 2019, we observed substantial support within the Yupik community for language revitalization and for resource development to support Yupik education. To that end, Chen \& Schwartz (2018) implemented a finite-state morphological analyzer as a critical enabling technology for use in Yupik language education and technology. Chen \& Schwartz (2018) reported a morphological analysis coverage rate of approximately 75\% on a dataset of 60K Yupik tokens, leaving considerable room for improvement. In this work, we present a re-implementation of the Chen \& Schwartz (2018) finite-state morphological analyzer for St. Lawrence Island Yupik that incorporates new linguistic insights; in particular, in this implementation we make use of the Paradigm Function Morphology (PFM) theory of morphology. We evaluate this new PFM-based morphological analyzer, and demonstrate that it consistently outperforms the existing analyzer of Chen \& Schwartz (2018) with respect to accuracy and coverage rate across multiple datasets.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.326}
}

@InProceedings{himoro-parejalora:2020:LREC,
  author   = {Himoro, Marcelo Yuji and Pareja-Lora, Antonio},
  title   = {Towards a Spell Checker for Zamboanga Chavacano Orthography},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2685--2697},
  abstract = {Zamboanga Chabacano (ZC) is the most vibrant variety of Philippine Creole Spanish, with over 400,000 native speakers in the Philippines (as of 2010). Following its introduction as a subject and a medium of instruction in the public schools of Zamboanga City from Grade 1 to 3 in 2012, an official orthography for this variety - the so-called “Zamboanga Chavacano Orthography” - has been approved in 2014. Its complexity, however, is a barrier to most speakers, since it does not necessarily reflect the particular phonetic evolution in ZC, but favours etymology instead. The distance between the correct spelling and the different spelling variations is often so great that delivering acceptable performance with the current de facto spell checking technologies may be challenging. The goals of this research have been to propose i) a spelling error taxonomy for ZC, formalised as an ontology and ii) an adaptive spell checking approach using Character-Based Statistical Machine Translation to correct spelling errors in ZC. Our results show that this approach is suitable for the goals mentioned and that it could be combined with other current spell checking technologies to achieve even higher performance.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.327}
}

```
@InProceedings{adouane-touileb-bernardy:2020:LREC,
  author    = {Adouane, Wafia  and  Touileb, Samia  and  Bernardy,
Jean-Philippe},
  title     = {Identifying Sentiments in Algerian Code-switched
User-generated Comments},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2698--2705},
  abstract  = {We present in this paper our work on Algerian
language, an under-resourced North African colloquial Arabic
variety, for which we built a comparably large corpus of more than
36,000 code-switched user-generated comments annotated for
sentiments. We opted for this data domain because Algerian is a
colloquial language with no existing freely available corpora.
Moreover, we compiled sentiment lexicons of positive and negative
unigrams and bigrams reflecting the code-switches present in the
language. We compare the performance of four models on the task of
identifying sentiments, and the results indicate that a CNN model
trained end-to-end fits better our unedited code-switched and
unbalanced data across the predefined sentiment classes.
Additionally, injecting the lexicons as background knowledge to the
model boosts its performance on the minority class with a gain of
10.54 points on the F-score. The results of our experiments can be
used as a baseline for future research for Algerian sentiment
analysis.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.328}
}

@InProceedings{linder-EtAl:2020:LREC,
  author    = {Linder, Lucy  and  Jungo, Michael  and  Hennebert,
Jean  and  Musat, Claudiu Cristian  and  Fischer, Andreas},
  title     = {Automatic Creation of Text Corpora for Low-Resource
Languages from the Internet: The Case of Swiss German},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2706--2711},
  abstract  = {This paper presents SwissCrawl, the largest Swiss
German text corpus to date. Composed of more than half a million
sentences, it was generated using a customized web scraping tool
that could be applied to other low-resource languages as well. The
approach demonstrates how freely available web pages can be used to
construct comprehensive text corpora, which are of fundamental
importance for natural language processing. In an experimental
evaluation, we show that using the new corpus leads to significant
improvements for the task of language modeling.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.329}
```

}

@InProceedings{chang-hsieh:2020:LREC,
  author   = {Chang, Yu-Yun and Hsieh, Shu-Kai},
  title   = {Do You Believe It Happened? Assessing Chinese Readers' Veridicality Judgments},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {259--267},
  abstract = {This work collects and studies Chinese readers' veridicality judgments to news events (whether an event is viewed as happening or not). For instance, in "The FBI alleged in court documents that Zazi had admitted having a handwritten recipe for explosives on his computer", do people believe that Zazi had a handwritten recipe for explosives? The goal is to observe the pragmatic behaviors of linguistic features under context which affects readers in making veridicality judgments. Exploring from the datasets, it is found that features such as event-selecting predicates (ESP), modality markers, adverbs, temporal information, and statistics have an impact on readers' veridicality judgments. We further investigated that modality markers with high certainty do not necessarily trigger readers to have high confidence in believing an event happened. Additionally, the source of information introduced by an ESP presents low effects to veridicality judgments, even when an event is attributed to an authority (e.g. "The FBI"). A corpus annotated with Chinese readers' veridicality judgments is released as the Chinese PragBank for further analysis.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.33}
}

@InProceedings{hakimiparizi-cook:2020:LREC,
  author   = {Hakimi Parizi, Ali and Cook, Paul},
  title   = {Evaluating Sub-word Embeddings in Cross-lingual Models},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2712--2719},
  abstract = {Cross-lingual word embeddings create a shared space for embeddings in two languages, and enable knowledge to be transferred between languages for tasks such as bilingual lexicon induction. One problem, however, is out-of-vocabulary (OOV) words, for which no embeddings are available. This is particularly problematic for low-resource and morphologically-rich languages, which often have relatively high OOV rates. Approaches to learning sub-word embeddings have been proposed to address the problem of OOV words, but most prior work has not considered sub-word embeddings in

```
cross-lingual models. In this paper, we consider whether sub-word
embeddings can be leveraged to form cross-lingual embeddings for OOV
words. Specifically, we consider a novel bilingual lexicon induction
task focused on OOV words, for language pairs covering several
language families. Our results indicate that cross-lingual
representations for OOV words can indeed be formed from sub-word
embeddings, including in the case of a truly low-resource
morphologically-rich language.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.330}
}

@InProceedings{schmidt-EtAl:2020:LREC,
  author    = {Schmidt, Larissa  and  Linder, Lucy  and
Djambazovska, Sandra  and  Lazaridis, Alexandros  and  Samardžić,
Tanja  and  Musat, Claudiu},
  title     = {A Swiss German Dictionary: Variation in Speech and
Writing},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2720--2725},
  abstract  = {We introduce a dictionary containing normalized forms
of common words in various Swiss German dialects into High German.
As Swiss German is, for now, a predominantly spoken language, there
is a significant variation in the written forms, even between
speakers of the same dialect. To alleviate the uncertainty
associated with this diversity, we complement the pairs of Swiss
German - High German words with the Swiss German phonetic
transcriptions (SAMPA). This dictionary becomes thus the first
resource to combine large-scale spontaneous translation with
phonetic transcriptions. Moreover, we control for the regional
distribution and insure the equal representation of the major Swiss
dialects. The coupling of the phonetic and written Swiss German
forms is powerful. We show that they are sufficient to train a
Transformer-based phoneme to grapheme model that generates credible
novel Swiss German writings. In addition, we show that the inverse
mapping - from graphemes to phonemes - can be modeled with a
transformer trained with the novel dictionary. This generation of
pronunciations for previously unknown words is key in training
extensible automated speech recognition (ASR) systems, which are key
beneficiaries of this dictionary.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.331}
}

@InProceedings{kevers-retalimedori:2020:LREC,
  author    = {Kevers, Laurent  and  Retali-Medori, Stella},
  title     = {Towards a Corsican Basic Language Resource Kit},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2726--2735},
  abstract  = {The current situation regarding the existence of
natural language processing (NLP) resources and tools for Corsican
reveals their virtual non-existence. Our inventory contains only a
few rare digital resources, lexical or corpus databases, requiring
adaptation work. Our objective is to use the Banque de Données
Langue Corse project (BDLC) to improve the availability of resources
and tools for the Corsican language and, in the long term, provide a
complete Basic Language Ressource Kit (BLARK). We have defined a
roadmap setting out the actions to be undertaken: the collection of
corpora and the setting up of a consultation interface
(concordancer), and of a language detection tool, an electronic
dictionary and a part-of-speech tagger. The first achievements
regarding these topics have already been reached and are presented
in this article. Some elements are also available on our project
page (http://bdlc.univ-corse.fr/tal/).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.332}
}

@InProceedings{boudreau-EtAl:2020:LREC,
  author    = {Boudreau, Jeremie  and  Patra, Akankshya  and
Suvarna, Ashima  and  Cook, Paul},
  title     = {Evaluating the Impact of Sub-word Information and
Cross-lingual Word Embeddings on Mi'kmaq Language Modelling},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2736--2745},
  abstract  = {Mi'kmaq is an Indigenous language spoken primarily in
Eastern Canada. It is polysynthetic and low-resource. In this paper
we consider a range of n-gram and RNN language models for Mi'kmaq.
We find that an RNN language model, initialized with pre-trained
fastText embeddings, performs best, highlighting the importance of
sub-word information for Mi'kmaq language modelling. We further
consider approaches to language modelling that incorporate cross-
lingual word embeddings, but do not see improvements with these
models. Finally we consider language models that operate over
segmentations produced by SentencePiece --- which include sub-word
units as tokens --- as opposed to word-level models. We see
improvements for this approach over word-level language models,
again indicating that sub-word modelling is important for Mi'kmaq
language modelling.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.333}
}

@InProceedings{brixey-EtAl:2020:LREC,
  author    = {Brixey, Jacqueline  and  Sides, David  and  Vizthum,
Timothy  and  Traum, David  and  Iskarous, Khalil},
  title     = {Exploring a Choctaw Language Corpus with Word Vectors
```

and Minimum Distance Length},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2746--2753},
  abstract = {This work introduces additions to the corpus ChoCo, a multimodal corpus for the American indigenous language Choctaw. Using texts from the corpus, we develop new computational resources by using two off-the-shelf tools: word2vec and Linguistica. Our work illustrates how these tools can be successfully implemented with a small corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.334}
}

@InProceedings{alabi-EtAl:2020:LREC,
  author   = {Alabi, Jesujoba and Amponsah-Kaakyire, Kwabena and Adelani, David and España-Bonet, Cristina},
  title   = {Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2754--2762},
  abstract = {The success of several architectures to learn semantic representations from unannotated text and the availability of these kind of texts in online multilingual resources such as Wikipedia has facilitated the massive and automatic creation of resources for multiple languages. The evaluation of such resources is usually done for the high-resourced languages, where one has a smorgasbord of tasks and test sets to evaluate on. For low-resourced languages, the evaluation is more difficult and normally ignored, with the hope that the impressive capability of deep learning architectures to learn (multilingual) representations in the high-resourced setting holds in the low-resourced setting too. In this paper we focus on two African languages, Yorùbá and Twi, and compare the word embeddings obtained in this way, with word embeddings obtained from curated corpora and a language-dependent processing. We analyse the noise in the publicly available corpora, collect high quality and noisy data for the two languages and quantify the improvements that depend not only on the amount of data but on the quality too. We also use different architectures that learn word representations both from surface forms and characters to further exploit all the available information which showed to be important for these languages. For the evaluation, we manually translate the wordsim-353 word pairs dataset from English into Yorùbá and Twi. We extend the analysis to contextual word embeddings and evaluate multilingual BERT on a named entity recognition task. For this, we annotate with named entities the Global Voices corpus for Yorùbá. As

output of the work, we provide corpora, embeddings and the test suits for both languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.335}
}

@InProceedings{kara-EtAl:2020:LREC,
  author   = {Kara, Neslihan and Aslan, Deniz Baran and Marşan, Büşra and Bakay, Özge and Ak, Koray and Yıldız, Olcay Taner},
  title   = {TRopBank: Turkish PropBank V2.0},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2763--2772},
  abstract = {In this paper, we present and explain TRopBank “Turkish PropBank v2.0”. PropBank is a hand-annotated corpus of propositions which is used to obtain the predicate-argument information of a language. Predicate-argument information of a language can help understand semantic roles of arguments. “Turkish PropBank v2.0”, unlike PropBank v1.0, has a much more extensive list of Turkish verbs, with 17.673 verbs in total.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.336}
}

@InProceedings{tufi-EtAl:2020:LREC,
  author   = {Tufiș, Dan and Mitrofan, Maria and Păiș, Vasile and Ion, Radu and Coman, Andrei},
  title   = {Collection and Annotation of the Romanian Legal Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2773--2777},
  abstract = {We present the Romanian legislative corpus which is a valuable linguistic asset for the development of machine translation systems, especially for under-resourced languages. The knowledge that can be extracted from this resource is necessary for a deeper understanding of how law terminology is used and how it can be made more consistent. At this moment the corpus contains more than 140k documents representing the legislative body of Romania. This corpus is processed and annotated at different levels: linguistically (tokenized, lemmatized and pos-tagged), dependency parsed, chunked, named entities identified and labeled with IATE terms and EUROVOC descriptors. Each annotated document has a CONLL-U Plus format consisting in 14 columns, in addition to the standard 10-column format, four other types of annotations were added. Moreover the repository will be periodically updated as new legislative texts are published. These will be automatically collected and transmitted to the processing and annotation pipeline. The access to the corpus

will be done through ELRC infrastructure.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.337}
}

@InProceedings{vonprince-nordhoff:2020:LREC,
  author   = {von Prince, Kilu and Nordhoff, Sebastian},
  title   = {An Empirical Evaluation of Annotation Practices in Corpora from Language Documentation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2778--2787},
  abstract = {For most of the world's languages, no primary data are available, even as many languages are disappearing. Throughout the last two decades, however, language documentation projects have produced substantial amounts of primary data from a wide variety of endangered languages. These resources are still in the early days of their exploration. One of the factors that makes them hard to use is a relative lack of standardized annotation conventions. In this paper, we will describe common practices in existing corpora in order to facilitate their future processing. After a brief introduction of the main formats used for annotation files, we will focus on commonly used tiers in the widespread ELAN and Toolbox formats. Minimally, corpora from language documentation contain a transcription tier and an aligned translation tier, which means they constitute parallel corpora. Additional common annotations include named references, morpheme separation, morpheme-by-morpheme glosses, part-of-speech tags and notes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.338}
}

@InProceedings{mohanty-mishra-mamidi:2020:LREC,
  author   = {Mohanty, Gaurav and Mishra, Pruthwik and Mamidi, Radhika},
  title   = {Annotated Corpus for Sentiment Analysis in Odia Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2788--2795},
  abstract = {Given the lack of an annotated corpus of non-traditional Odia literature which serves as the standard when it comes sentiment analysis, we have created an annotated corpus of Odia sentences and made it publicly available to promote research in the field. Secondly, in order to test the usability of currently available Odia sentiment lexicon, we experimented with various classifiers by training and testing on the sentiment annotated corpus while using identified affective words from the same as

features. Annotation and classification are done at sentence level as the usage of sentiment lexicon is best suited to sentiment analysis at this level. The created corpus contains 2045 Odia sentences from news domain annotated with sentiment labels using a well-defined annotation scheme. An inter-annotator agreement score of 0.79 is reported for the corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.339}
}

@InProceedings{nicolas-EtAl:2020:LREC,
  author   = {Nicolas, Lionel and Lyding, Verena and Borg, Claudia and Forascu, Corina and Fort, Karën and Zdravkova, Katerina and Kosem, Iztok and Čibej, Jaka and Arhar Holdt, Špela and Millour, Alice and König, Alexander and Rodosthenous, Christos and Sangati, Federico and ul Hassan, Umair and Katinskaia, Anisia and Barreiro, Anabela and Aparaschivei, Lavinia and HaCohen-Kerner, Yaakov},
  title   = {Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {268--278},
  abstract = {We introduce in this paper a generic approach to combine implicit crowdsourcing and language learning in order to mass-produce language resources (LRs) for any language for which a crowd of language learners can be involved. We present the approach by explaining its core paradigm that consists in pairing specific types of LRs with specific exercises, by detailing both its strengths and challenges, and by discussing how much these challenges have been addressed at present. Accordingly, we also report on on-going proof-of-concept efforts aiming at developing the first prototypical implementation of the approach in order to correct and extend an LR called ConceptNet based on the input crowdsourced from language learners. We then present an international network called the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect) that provides the context to accelerate the implementation of this generic approach. Finally, we exemplify how it can be used in several language learning scenarios to produce a multitude of NLP resources and how it can therefore alleviate the long-standing NLP issue of the lack of LRs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.34}
}

@InProceedings{lpezdelacalle-saralegi-sanvicente:2020:LREC,
  author   = {López de Lacalle, Maddalen and Saralegi, Xabier and San Vicente, Iñaki},
  title   = {Building a Task-oriented Dialog System for Languages with no Training Data: the Case for Basque},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2796--2802},
  abstract  = {This paper presents an approach for developing a
task-oriented dialog system for less-resourced languages in
scenarios where training data is not available. Both intent
classification and slot filling are tackled. We project the existing
annotations in rich-resource languages by means of Neural Machine
Translation (NMT) and posterior word alignments. We then compare
training on the projected monolingual data with direct model
transfer alternatives. Intent Classifiers and slot filling sequence
taggers are implemented using a BiLSTM architecture or by fine-
tuning BERT transformer models. Models learnt exclusively from
Basque projected data provide better accuracies for slot filling.
Combining Basque projected train data with rich-resource languages
data outperforms consistently models trained solely on projected
data for intent classification. At any rate, we achieve competitive
performance in both tasks, with accuracies of 81\% for intent
classification and 77\% for slot filling.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.340}
}

@InProceedings{nguer-EtAl:2020:LREC,
  author    = {Nguer, Elhadji Mamadou  and  Lo, Alla  and  Dione,
Cheikh M. Bamba  and  Ba, Sileye O.  and  Lo, Moussa},
  title     = {SENCORPUS: A French-Wolof Parallel Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2803--2811},
  abstract  = {In this paper, we report efforts towards the
acquisition and construction of a bilingual parallel corpus between
French and Wolof, a Niger-Congo language belonging to the Northern
branch of the Atlantic group. The corpus is constructed as part of
the SYSNET3LOc project. It currently contains about 70,000 French-
Wolof parallel sentences drawn on various sources from different
domains. The paper discusses the data collection procedure,
conversion, and alignment of the corpus as well as it’s application
as training data for neural machine translation. In fact, using this
corpus, we were able to create word embedding models for Wolof with
relatively good results. Currently, the corpus is being used to
develop a neural machine translation model to translate French
sentences into Wolof.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.341}
}

@InProceedings{bella-EtAl:2020:LREC,
```

```
  author    = {Bella, Gábor  and  McNeill, Fiona  and  Gorman, Rody
and  O Donnaile, Caoimhin  and  MacDonald, Kirsty  and
Chandrashekar, Yamini  and  Freihat, Abed Alhakim  and  Giunchiglia,
Fausto},
  title     = {A Major Wordnet for a Minority Language: Scottish
Gaelic},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2812--2818},
  abstract  = {We present a new wordnet resource for Scottish
Gaelic, a Celtic minority language spoken by about 60,000 speakers,
most of whom live in Northwestern Scotland. The wordnet contains
over 15 thousand word senses and was constructed by merging ten
thousand new, high-quality translations, provided and validated by
language experts, with an existing wordnet derived from Wiktionary.
This new, considerably extended wordnet—currently among the 30
largest in the world—targets multiple communities: language speakers
and learners; linguists; computer scientists solving problems
related to natural language processing. By publishing it as a freely
downloadable resource, we hope to contribute to the long-term
preservation of Scottish Gaelic as a living language, both offline
and on the Web.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.342}
}

@InProceedings{abraham-EtAl:2020:LREC,
  author    = {Abraham, Basil  and  Goel, Danish  and  Siddarth,
Divya  and  Bali, Kalika  and  Chopra, Manu  and  Choudhury, Monojit
and  Joshi, Pratik  and  Jyoti, Preethi  and  Sitaram, Sunayana  and
Seshadri, Vivek},
  title     = {Crowdsourcing Speech Data for Low-Resource Languages
from Low-Income Workers},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2819--2826},
  abstract  = {Voice-based technologies are essential to cater to
the hundreds of millions of new smartphone users. However, most of
the languages spoken by these new users have little to no labelled
speech data. Unfortunately, collecting labelled speech data in any
language is an expensive and resource-intensive task. Moreover,
existing platforms typically collect speech data only from urban
speakers familiar with digital technology whose dialects are often
very different from low-income users. In this paper, we explore the
possibility of collecting labelled speech data directly from low-
income workers. In addition to providing diversity to the speech
dataset, we believe this approach can also provide valuable
```

```
supplemental earning opportunities to these communities. To this
end, we conducted a study where we collected labelled speech data in
the Marathi language from three different user groups: low-income
rural users, low-income urban users, and university students.
Overall, we collected 109 hours of data from 36 participants. Our
results show that the data collected from low-income participants is
of comparable quality to the data collected from university students
(who are typically employed to do this work) and that crowdsourcing
speech data from low-income rural and urban workers is a viable
method of gathering speech data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.343}
}

@InProceedings{cruz-anastasopoulos-stump:2020:LREC,
  author    = {Cruz, Hilaria  and  Anastasopoulos, Antonios  and
Stump, Gregory},
  title     = {A Resource for Studying Chatino Verbal Morphology},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2827--2831},
  abstract  = {We present the first resource focusing on the verbal
inflectional morphology of San Juan Quiahije Chatino, a tonal
mesoamerican language spoken in Mexico. We provide a collection of
complete inflection tables of 198 lemmata, with morphological tags
based on the UniMorph schema. We also provide baseline results on
three core NLP tasks: morphological analysis, lemmatization, and
morphological inflection.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.344}
}

@InProceedings{mehta-EtAl:2020:LREC,
  author    = {Mehta, Devansh  and  Santy, Sebastin  and  Mothilal,
Ramaravind Kommiya  and  Srivastava, Brij Mohan Lal  and  Sharma,
Alok  and  Shukla, Anurag  and  Prasad, Vishnu  and  U, Venkanna
and  Sharma, Amit  and  Bali, Kalika},
  title     = {Learnings from Technological Interventions in a Low
Resource Language: A Case-Study on Gondi},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2832--2838},
  abstract  = {The primary obstacle to developing technologies for
low-resource languages is the lack of usable data. In this paper, we
report the adaption and deployment of 4 technology-driven methods of
data collection for Gondi, a low-resource vulnerable language spoken
by around 2.3 million tribal people in south and central India. In
the process of data collection, we also help in its revival by
```

expanding access to information in Gondi through the creation of linguistic resources that can be used by the community, such as a dictionary, children's stories, an app with Gondi content from multiple sources and an Interactive Voice Response (IVR) based mass awareness platform. At the end of these interventions, we collected a little less than 12,000 translated words and/or sentences and identified more than 650 community members whose help can be solicited for future translation efforts. The larger goal of the project is collecting enough data in Gondi to build and deploy viable language technologies like machine translation and speech to text systems that can help take the language onto the internet.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.345}
}

@InProceedings{golazizian-EtAl:2020:LREC,
  author   = {Golazizian, Preni and Sabeti, Behnam and Ashrafi Asli, Seyed Arad and Majdabadi, Zahra and Momenzadeh, Omid and fahmi, reza},
  title   = {Irony Detection in Persian Language: A Transfer Learning Approach Using Emoji Prediction},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2839--2845},
  abstract = {Irony is a linguistic device used to intend an idea while articulating an opposing expression. Many text analytic algorithms used for emotion extraction or sentiment analysis, produce invalid results due to the use of irony. Persian speakers use this device more often due to the language’s nature and some cultural reasons. This phenomenon also appears in social media platforms such as Twitter where users express their opinions using ironic or sarcastic posts. In the current research, which is the first attempt at irony detection in Persian language, emoji prediction is used to build a pretrained model. The model is finetuned utilizing a set of hand labeled tweets with irony tags. A bidirectional LSTM (BiLSTM) network is employed as the basis of our model which is improved by attention mechanism. Additionally, a Persian corpus for irony detection containing 4339 manually-labeled tweets is introduced. Experiments show the proposed approach outperforms the adapted state-of-the-art method tested on Persian dataset with an accuracy of 83.1\%, and offers a strong baseline for further research in Persian language.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.346}
}

@InProceedings{bamutura-ljunglf-nebende:2020:LREC,
  author   = {Bamutura, David and Ljunglöf, Peter and Nebende, Peter},
  title   = {Towards Computational Resource Grammars for Runyankore and Rukiga},
  booktitle   = {Proceedings of The 12th Language Resources and

Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2846--2854},
  abstract = {In this paper, we present computational resource grammars of Runyankore and Rukiga (R\&R) languages. Runyankore and Rukiga are two under-resourced Bantu Languages spoken by about 6 million people indigenous to South- Western Uganda, East Africa. We used Grammatical Framework (GF), a multilingual grammar formalism and a special- purpose functional programming language to formalise the descriptive grammar of these languages. To the best of our knowledge, these computational resource grammars are the first attempt to the creation of language resources for R\&R. In Future Work, we plan to use these grammars to bootstrap the generation of other linguistic resources such as multilingual corpora that make use of data-driven approaches to natural language processing feasible. In the meantime, they can be used to build Computer-Assisted Language Learning (CALL) applications for these languages among others.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.347}
}

@InProceedings{ashrafiasli-EtAl:2020:LREC,
  author   = {Ashrafi Asli, Seyed Arad and Sabeti, Behnam and Majdabadi, Zahra and Golazizian, Preni and fahmi, reza and Momenzadeh, Omid},
  title   = {Optimizing Annotation Effort Using Active Learning Strategies: A Sentiment Analysis Case Study in Persian},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2855--2861},
  abstract = {Deep learning models are the current State-of-the-art methodologies towards many real-world problems. However, they need a substantial amount of labeled data to be trained appropriately. Acquiring labeled data can be challenging in some particular domains or less-resourced languages. There are some practical solutions regarding these issues, such as Active Learning and Transfer Learning. Active learning's idea is simple: let the model choose the samples for annotation instead of labeling the whole dataset. This method leads to a more efficient annotation process. Active Learning models can achieve the baseline performance (the accuracy of the model trained on the whole dataset), with a considerably lower amount of labeled data. Several active learning approaches are tested in this work, and their compatibility with Persian is examined using a brand-new sentiment analysis dataset that is also introduced in this work. MirasOpinion, which to our knowledge is the largest Persian sentiment analysis dataset, is crawled from a Persian e-commerce website and annotated using a crowd-sourcing

policy. LDA sampling, which is an efficient Active Learning strategy using Topic Modeling, is proposed in this research. Active Learning Strategies have shown promising results in the Persian language, and LDA sampling showed a competitive performance compared to other approaches.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.348}
}

@InProceedings{hossain-EtAl:2020:LREC,
  author   = {Hossain, Md Zobaer and Rahman, Md Ashraful and Islam, Md Saiful and Kar, Sudipta},
  title   = {BanFakeNews: A Dataset for Detecting Fake News in Bangla},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2862--2871},
  abstract = {Observing the damages that can be done by the rapid propagation of fake news in various sectors like politics and finance, automatic identification of fake news using linguistic analysis has drawn the attention of the research community. However, such methods are largely being developed for English where low resource languages remain out of the focus. But the risks spawned by fake and manipulative news are not confined by languages. In this work, we propose an annotated dataset of ≈ 50K news that can be used for building automated fake news detection systems for a low resource language like Bangla. Additionally, we provide an analysis of the dataset and develop a benchmark system with state of the art NLP techniques to identify Bangla fake news. To create this system, we explore traditional linguistic features and neural network based methods. We expect this dataset will be a valuable resource for building technologies to prevent the spreading of fake news and contribute in research with low resource languages. The dataset and source code are publicly available at https://github.com/Rowan1697/FakeNews.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.349}
}

@InProceedings{haagsma-bos-nissim:2020:LREC,
  author   = {Haagsma, Hessel and Bos, Johan and Nissim, Malvina},
  title   = {MAGPIE: A Large Corpus of Potentially Idiomatic Expressions},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {279--287},
  abstract = {Given the limited size of existing idiom corpora, we

```
aim to enable progress in automatic idiom processing and linguistic analysis by creating the largest-to-date corpus of idioms for English. Using a fixed idiom list, automatic pre-extraction, and a strictly controlled crowdsourced annotation procedure, we show that it is feasible to build a high-quality corpus comprising more than 50K instances, an order of a magnitude larger than previous resources. Crucial ingredients of crowdsourcing were the selection of crowdworkers, clear and comprehensive instructions, and an interface that breaks down the task in small, manageable steps. Analysis of the resulting corpus revealed strong effects of genre on idiom distribution, providing new evidence for existing theories on what influences idiom usage. The corpus also contains rich metadata, and is made publicly available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.35}
}

@InProceedings{duan-EtAl:2020:LREC,
  author    = {Duan, Mingjun  and  Fasola, Carlos  and  Rallabandi, Sai Krishna  and  Vega, Rodolfo  and  Anastasopoulos, Antonios  and  Levin, Lori  and  Black, Alan W},
  title     = {A Resource for Computational Experiments on Mapudungun},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2872--2877},
  abstract  = {We present a resource for computational experiments on Mapudungun, a polysynthetic indigenous language spoken in Chile with upwards of 200 thousand speakers. We provide 142 hours of culturally significant conversations in the domain of medical treatment. The conversations are fully transcribed and translated into Spanish. The transcriptions also include annotations for code-switching and non-standard pronunciations. We also provide baseline results on three core NLP tasks: speech recognition, speech synthesis, and machine translation between Spanish and Mapudungun. We further explore other applications for which the corpus will be suitable, including the study of code-switching, historical orthography change, linguistic structure, and sociological and anthropological studies.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.350}
}

@InProceedings{round-EtAl:2020:LREC,
  author    = {Round, Erich  and  Ellison, Mark  and  Macklin-Cordes, Jayden  and  Beniamine, Sacha},
  title     = {Automated Parsing of Interlinear Glossed Text from Page Images of Grammatical Descriptions},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2878--2883},
  abstract  = {Linguists seek insight from all human languages,
however accessing information from most of the full store of extant
global linguistic descriptions is not easy. One of the most common
kinds of information that linguists have documented is vernacular
sentences, as recorded in descriptive grammars. Typically these
sentences are formatted as interlinear glossed text (IGT). Most
descriptive grammars, however, exist only as hardcopy or scanned pdf
documents. Consequently, parsing IGTs in scanned grammars is a
priority, in order to significantly increase the volume of
documented linguistic information that is readily accessible. Here
we demonstrate fundamental viability for a technology that can
assist in making a large number of linguistic data sources machine
readable: the automated identification and parsing of interlinear
glossed text from scanned page images. For example, we attain high
median precision and recall (>0.95) in the identification of
examples sentences in IGT format. Our results will be of interest to
those who are keen to see more of the existing documentation of
human language, especially for less-resourced and endangered
languages, become more readily accessible.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.351}
}

@InProceedings{mccarthy-EtAl:2020:LREC1,
  author    = {McCarthy, Arya D.  and  Wicks, Rachel  and  Lewis,
Dylan  and  Mueller, Aaron  and  Wu, Winston  and  Adams, Oliver
and  Nicolai, Garrett  and  Post, Matt  and  Yarowsky, David},
  title     = {The Johns Hopkins University Bible Corpus: 1600+
Tongues for Typological Exploration},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2884--2892},
  abstract  = {We present findings from the creation of a massively
parallel corpus in over 1600 languages, the Johns Hopkins University
Bible Corpus (JHUBC). The corpus consists of over 4000 unique
translations of the Christian Bible and counting. Our data is
derived from scraping several online resources and merging them with
existing corpora, combining them under a common scheme that is
verse-parallel across all translations. We detail our effort to
scrape, clean, align, and utilize this ripe multilingual dataset.
The corpus captures the great typological variety of the world's
languages. We catalog this by showing highly similar proportions of
representation of Ethnologue's typological features in our corpus.
We also give an example application: projecting pronoun features
like clusivity across alignments to richly annotate languages which
do not mark the distinction.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.352}
}
```

```
@InProceedings{zahrer-zgank-schuppler:2020:LREC,
  author    = {Zahrer, Alexander  and  Zgank, Andrej  and
Schuppler, Barbara},
  title     = {Towards Building an Automatic Transcription System
for Language Documentation: Experiences from Muyu},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2893--2900},
  abstract  = {Since at least half of the world’s 6000 plus
languages will vanish during the 21st century, language
documentation has become a rapidly growing field in linguistics. A
fundamental challenge for language documentation is the
”transcription bottleneck”. Speech technology may deliver the
decisive breakthrough for overcoming the transcription bottleneck.
This paper presents first experiments from the development of
ASR4LD, a new automatic speech recognition (ASR) based tool for
language documentation (LD). The experiments are based on recordings
from an ongoing documentation project for the endangered Muyu
language in New Guinea. We compare phoneme recognition experiments
with American English, Austrian German and Slovenian as source
language and Muyu as target language. The Slovenian acoustic models
achieve the by far best performance (43.71\% PER) in comparison to
57.14\% PER with American English, and 89.49\% PER with Austrian
German. Whereas part of the errors can be explained by phonetic
variation, the recording mismatch poses a major problem. On the long
term, ASR4LD will not only be an integral part of the ongoing
documentation project of Muyu, but will be further developed in
order to facilitate also the language documentation process of other
language groups.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.353}
}

@InProceedings{jettka-lehmberg:2020:LREC,
  author    = {Jettka, Daniel  and  Lehmberg, Timm},
  title     = {Towards Flexible Cross-Resource Exploitation of
Heterogeneous Language Documentation Data},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2901--2905},
  abstract  = {This paper reports on challenges and solution
approaches in the development of methods for language resource
overarching data analysis in the field of language documentation. It
is based on the successful outcomes of the initial phase of an 18
year long-term project on lesser resourced and mostly endangered
indigenous languages of the Northern Eurasian area, which included
```

```
the finalization and publication of multiple language corpora and additional language resources. While aiming at comprehensive cross-resource data analysis, the project at the same time is confronted with a dynamic and complex resource landscape, especially resulting from a vast amount of multi-layered information stored in the form of analogue primary data in different widespread archives on the territory of the Russian Federation. The methods described aim at solving the tension between unification of data sets and vocabularies on the one hand and maximum openness for the integration of future resources and adaption of external information on the other hand.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.354}
}

@InProceedings{winterstein-tang-lai:2020:LREC,
  author    = {Winterstein, Grégoire  and  Tang, Carmen  and  Lai, Regine},
  title     = {CantoMap: a Hong Kong Cantonese MapTask Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2906--2913},
  abstract  = {This work reports on the construction of a corpus of connected spoken Hong Kong Cantonese. The corpus aims at providing an additional resource for the study of modern (Hong Kong) Cantonese and also involves several controlled elicitation tasks which will serve different projects related to the phonology and semantics of Cantonese. The word-segmented corpus offers recordings, phonemic transcription, and Chinese characters transcription. The corpus contains a total of 768 minutes of recordings and transcripts of forty speakers. All the audio material has been aligned at utterance level with the transcriptions, using the ELAN transcription and annotation tool. The controlled elicitation task was based on the design of HCRC MapTask corpus (Anderson et al., 1991), in which participants had to communicate using solely verbal means as eye contact was restricted. In this paper, we outline the design of the maps and their landmarks and the basic segmentation principles of the data and various transcription conventions we adopted. We also compare the contents of Cantomap to those of comparable Cantonese corpora.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.355}
}

@InProceedings{bustamante-oncevay-zariquiey:2020:LREC,
  author    = {Bustamante, Gina  and  Oncevay, Arturo  and  Zariquiey, Roberto},
  title     = {No Data to Crawl? Monolingual Corpus Creation from PDF Files of Truly low-Resource Languages in Peru},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {2914--2923},
  abstract  = {We introduce new monolingual corpora for four
indigenous and endangered languages from Peru: Shipibo-konibo,
Ashaninka, Yanesha and Yine. Given the total absence of these
languages in the web, the extraction and processing of texts from
PDF files is relevant in a truly low-resource language scenario. Our
procedure for monolingual corpus creation considers language-
specific and language-agnostic steps, and focuses on educational PDF
files with multilingual sentences, noisy pages and low-structured
content. Through an evaluation based on language modelling and
character-level perplexity on a subset of manually extracted
sentences, we determine that our method allows the creation of clean
corpora for the four languages, a key resource for natural language
processing tasks nowadays.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.356}
}

@InProceedings{jnsdttir-ingason:2020:LREC,
  author    = {Jónsdóttir, Hildur  and  Ingason, Anton Karl},
  title     = {Creating a Parallel Icelandic Dependency Treebank
from Raw Text to Universal Dependencies},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2924--2931},
  abstract  = {Making the low-resource language, Icelandic,
accessible and usable in Language Technology is a work in progress
and is supported by the Icelandic government. Creating resources and
suitable training data (e.g., a dependency treebank) is a
fundamental part of that work. We describe work on a parallel
Icelandic dependency treebank based on Universal Dependencies (UD).
This is important because it is the first parallel treebank resource
for the language and since several other languages already have a
resource based on the same text. Two Icelandic treebanks based on
phrase-structure grammar have been built and ongoing work aims to
convert them to UD. Previously, limited work has been done on
dependency grammar for Icelandic. The current project aims to
ameliorate this situation by creating a small dependency treebank
from scratch. Creating a treebank is a laborious task so the process
was implemented in an accessible manner using freely available tools
and resources. The parallel data in the UD project was chosen as a
source because this would furthermore give us the first parallel
treebank for Icelandic. The Icelandic parallel UD corpus will be
published as part of UD version 2.6.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.357}
}

@InProceedings{miletic-EtAl:2020:LREC,
```

```
  author    = {Miletic, Aleksandra  and  Bras, Myriam  and  Vergez-
Couret, Marianne  and  Esher, Louise  and  Poujade, Clamença  and
Sibille, Jean},
  title     = {Building a Universal Dependencies Treebank for
Occitan},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2932--2939},
  abstract  = {This paper outlines the ongoing effort of creating
the first treebank for Occitan, a low-ressourced regional language
spoken mainly in the south of France. We briefly present the global
context of the project and report on its current status. We adopt
the Universal Dependencies framework for this project. Our
methodology is based on two main principles. Firstly, in order to
guarantee the annotation quality, we use the agile annotation
approach. Secondly, we rely on pre-processing using existing tools
(taggers and parsers) to facilitate the work of human annotators,
mainly through a delexicalized cross-lingual parsing approach. We
present the results available at this point (annotation guidelines
and a sub-corpus annotated with PoS tags and lemmas) and give the
timeline for the rest of the work.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.358}
}

@InProceedings{moeljadi-aminullah:2020:LREC,
  author    = {Moeljadi, David  and  Aminullah, Zakariya Pamuji},
  title     = {Building the Old Javanese Wordnet},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2940--2946},
  abstract  = {This paper discusses the construction and the ongoing
development of the Old Javanese Wordnet. The words were extracted
from the digitized version of the Old Javanese–English Dictionary
(Zoetmulder, 1982). The wordnet is built using the 'expansion'
approach (Vossen, 1998), leveraging on the Princeton Wordnet's core
synsets and semantic hierarchy, as well as scientific names. The
main goal of our project was to produce a high quality, human-
curated resource. As of December 2019, the Old Javanese Wordnet
contains 2,054 concepts or synsets and 5,911 senses. It is released
under a Creative Commons Attribution 4.0 International License (CC
BY 4.0). We are still developing it and adding more synsets and
senses. We believe that the lexical data made available by this
wordnet will be useful for a variety of future uses such as the
development of Modern Javanese Wordnet and many language processing
tasks and linguistic research on Javanese.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.359}
```

}

@InProceedings{chiyahgarcia-EtAl:2020:LREC,
  author   = {Chiyah Garcia, Francisco Javier and Lopes, José and Liu, Xingkun and Hastie, Helen},
  title   = {CRWIZ: A Framework for Crowdsourcing Real-Time Wizard-of-Oz Dialogues},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {288--297},
  abstract = {Large corpora of task-based and open-domain conversational dialogues are hugely valuable in the field of data-driven dialogue systems. Crowdsourcing platforms, such as Amazon Mechanical Turk, have been an effective method for collecting such large amounts of data. However, difficulties arise when task-based dialogues require expert domain knowledge or rapid access to domain-relevant information, such as databases for tourism. This will become even more prevalent as dialogue systems become increasingly ambitious, expanding into tasks with high levels of complexity that require collaboration and forward planning, such as in our domain of emergency response. In this paper, we propose CRWIZ: a framework for collecting real-time Wizard of Oz dialogues through crowdsourcing for collaborative, complex tasks. This framework uses semi-guided dialogue to avoid interactions that breach procedures and processes only known to experts, while enabling the capture of a wide variety of interactions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.36}
}

@InProceedings{sierramartnez-EtAl:2020:LREC,
  author   = {Sierra Martínez, Gerardo and Montaño, Cynthia and Bel-Enguix, Gemma and Córdova, Diego and Mota Montoya, Margarita},
  title   = {CPLM, a Parallel Corpus for Mexican Languages: Development and Interface},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2947--2952},
  abstract = {Mexico is a Spanish speaking country that has a great language diversity, with 68 linguistic groups and 364 varieties. As they face a lack of representation in education, government, public services and media, they present high levels of endangerment. Due to the lack of data available on social media and the internet, few technologies have been developed for these languages. To analyze different linguistic phenomena in the country, the Language Engineering Group developed the Corpus Paralelo de Lenguas Mexicanas

```
(CPLM) [The Mexican Languages Parallel Corpus], a collaborative
parallel corpus for the low-resourced languages of Mexico. The CPLM
aligns Spanish with six indigenous languages: Maya, Ch'ol, Mazatec,
Mixtec, Otomi, and Nahuatl. First, this paper describes the process
of building the CPLM: text searching, digitalization and alignment
process. Furthermore, we present some difficulties regarding
dialectal and orthographic variations. Second, we present the
interface and types of searching as well as the use of filters.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.360}
}

@InProceedings{ali-lu-xu:2020:LREC,
  author    = {Ali, Wazir  and  Lu, Junyu  and  Xu, Zenglin},
  title     = {SiNER: A Large Dataset for Sindhi Named Entity
Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2953--2961},
  abstract  = {We introduce the SiNER: a named entity recognition
(NER) dataset for low-resourced Sindhi language with quality
baselines. It contains 1,338 news articles and more than 1.35
million tokens collected from Kawish and Awami Awaz Sindhi
newspapers using the begin-inside-outside (BIO) tagging scheme. The
proposed dataset is likely to be a significant resource for
statistical Sindhi language processing. The ultimate goal of
developing SiNER is to present a gold-standard dataset for Sindhi
NER along with quality baselines. We implement several baseline
approaches of conditional random field (CRF) and recent popular
state-of-the-art bi-directional long-short term memory (Bi-LSTM)
models. The promising F1-score of 89.16 outputted by the Bi-LSTM-CRF
model with character-level representations demonstrates the quality
of our proposed SiNER dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.361}
}

@InProceedings{song-EtAl:2020:LREC2,
  author    = {Song, Li  and  Dai, Yuling  and  Liu, Yihuan  and
Li, Bin  and  Qu, Weiguang},
  title     = {Construct a Sense-Frame Aligned Predicate Lexicon for
Chinese AMR Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2962--2969},
  abstract  = {The study of predicate frame is an important topic
for semantic analysis. Abstract Meaning Representation (AMR) is an
emerging graph based semantic representation of a sentence. Since
```

core semantic roles defined in the predicate lexicon compose the backbone in an AMR graph, the construction of the lexicon becomes the key issue. The existing lexicons blur senses and frames of predicates, which needs to be refined to meet the tasks like word sense disambiguation and event extraction. This paper introduces the on-going project on constructing a novel predicate lexicon for Chinese AMR corpus. The new lexicon includes 14,389 senses and 10,800 frames of 8,470 words. As some senses can be aligned to more than one frame, and vice versa, we found the alignment between senses is not just one frame per sense. Explicit analysis is given for multiple aligned relations, which proves the necessity of the proposed lexicon for AMR corpus, and supplies real data for linguistic theoretical studies.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.362}
}

@InProceedings{han-jones-smeaton:2020:LREC,
  author   = {Han, Lifeng and Jones, Gareth and Smeaton, Alan},
  title   = {MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2970--2979},
  abstract = {Multi-word expressions (MWEs) are a hot topic in research in natural language processing (NLP), including topics such as MWE detection, MWE decomposition, and research investigating the exploitation of MWEs in other NLP fields such as Machine Translation. However, the availability of bilingual or multi-lingual MWE corpora is very limited. The only bilingual MWE corpora that we are aware of is from the PARSEME (PARSing and Multi-word Expressions) EU Project. This is a small collection of only 871 pairs of English-German MWEs. In this paper, we present multi-lingual and bilingual MWE corpora that we have extracted from root parallel corpora. Our collections are 3,159,226 and 143,042 bilingual MWE pairs for German-English and Chinese-English respectively after filtering. We examine the quality of these extracted bilingual MWEs in MT experiments. Our initial experiments applying MWEs in MT show improved translation performances on MWE terms in qualitative analysis and better general evaluation scores in quantitative analysis, on both German-English and Chinese-English language pairs. We follow a standard experimental pipeline to create our MultiMWE corpora which are available online. Researchers can use this free corpus for their own models or use them in a knowledge base as model features.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.363}
}

@InProceedings{myatmon-EtAl:2020:LREC,
  author   = {Myat Mon, Aye and Ding, Chenchen and Kaing, Hour and Mar Soe, Khin and Utiyama, Masao and Sumita, Eiichiro},

```
  title     = {A Myanmar (Burmese)-English Named Entity
Transliteration Dictionary},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2980--2983},
  abstract  = {Transliteration is generally a phonetically based
transcription across different writing systems. It is a crucial task
for various downstream natural language processing applications. For
the Myanmar (Burmese) language, robust automatic transliteration for
borrowed English words is a challenging task because of the complex
Myanmar writing system and the lack of data. In this study, we
constructed a Myanmar-English named entity dictionary containing
more than eighty thousand transliteration instances. The data have
been released under a CC BY-NC-SA license. We evaluated the
automatic transliteration performance using statistical and neural
network-based approaches based on the prepared data. The neural
network model outperformed the statistical model significantly in
terms of the BLEU score on the character level. Different units used
in the Myanmar script for processing were also compared and
discussed.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.364}
}

@InProceedings{li-yang-ma:2020:LREC,
  author    = {Li, Peng-Hsuan  and  Yang, Tsan-Yu  and  Ma, Wei-
Yun},
  title     = {CA-EHN: Commonsense Analogy from E-HowNet},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {2984--2990},
  abstract  = {Embedding commonsense knowledge is crucial for end-
to-end models to generalize inference beyond training corpora.
However, existing word analogy datasets have tended to be
handcrafted, involving permutations of hundreds of words with only
dozens of pre-defined relations, mostly morphological relations and
named entities. In this work, we model commonsense knowledge down to
word-level analogical reasoning by leveraging E-HowNet, an ontology
that annotates 88K Chinese words with their structured sense
definitions and English translations. We present CA-EHN, the first
commonsense word analogy dataset containing 90,505 analogies
covering 5,656 words and 763 relations. Experiments show that CA-EHN
stands out as a great indicator of how well word representations
embed commonsense knowledge. The dataset is publicly available at
\url{https://github.com/ckiplab/CA-EHN}.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.365}
}
```

@InProceedings{leone-EtAl:2020:LREC,
  author   = {Leone, Valentina and Siragusa, Giovanni and Di Caro, Luigi and Navigli, Roberto},
  title   = {Building Semantic Grams of Human Knowledge},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2991--3000},
  abstract = {Word senses are typically defined with textual definitions for human consumption and, in computational lexicons, put in context via lexical-semantic relations such as synonymy, antonymy, hypernymy, etc. In this paper we embrace a radically different paradigm that provides a slot-filler structure, called “semagram”, to define the meaning of words in terms of their prototypical semantic information. We propose a semagram-based knowledge model composed of 26 semantic relationships which integrates features from a range of different sources, such as computational lexicons and property norms. We describe an annotation exercise regarding 50 concepts over 10 different categories and put forward different automated approaches for extending the semagram base to thousands of concepts. We finally evaluated the impact of the proposed resource on a semantic similarity task, showing significant improvements over state-of-the-art word embeddings.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.366}
}

@InProceedings{uban-dinu:2020:LREC,
  author   = {Uban, Ana Sabina and Dinu, Liviu P.},
  title   = {Automatically Building a Multilingual Lexicon of False Friends With No Supervision},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3001--3007},
  abstract = {Cognate words, defined as words in different languages which derive from a common etymon, can be useful for language learners, who can leverage the orthographical similarity of cognates to more easily understand a text in a foreign language. Deceptive cognates, or false friends, do not share the same meaning anymore; these can be instead deceiving and detrimental for language acquisition or text understanding in a foreign language. We use an automatic method of detecting false friends from a set of cognates, in a fully unsupervised fashion, based on cross-lingual word embeddings. We implement our method for English and five Romance languages, including a low-resource language (Romanian), and evaluate it against two different gold standards. The method can be extended easily to any language pair, requiring only large

```
monolingual corpora for the involved languages and a small bilingual
dictionary for the pair. We additionally propose a measure of
"falseness" of a false friends pair. We publish freely the database
of false friends in the six languages, along with the falseness
scores for each cognate pair. The resource is the largest of the
kind that we are aware of, both in terms of languages covered and
number of word pairs.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.367}
}

@InProceedings{angelov:2020:LREC,
  author    = {Angelov, Krasimir},
  title     = {A Parallel WordNet for English, Swedish and
Bulgarian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3008--3015},
  abstract  = {We present the parallel creation of a WordNet
resource for Swedish and Bulgarian which is tightly aligned with the
Princeton WordNet. The alignment is not only on the synset level,
but also on word level, by matching words with their closest
translations in each language. We argue that the tighter alignment
is essential in machine translation and natural language generation.
About one-fifth of the lexical entries are also linked to the
corresponding Wikipedia articles. In addition to the traditional
semantic relations in WordNet, we also integrate morphological and
morpho-syntactic information. The resource comes with a corpus where
examples from Princeton WordNet are translated to Swedish and
Bulgarian. The examples are aligned on word and phrase level. The
new resource is open-source and in its development we used only
existing open-source resources.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.368}
}

@InProceedings{sajous-calderone-hathout:2020:LREC,
  author    = {Sajous, Franck  and  Calderone, Basilio  and
Hathout, Nabil},
  title     = {ENGLAWI: From Human- to Machine-Readable Wiktionary},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3016--3026},
  abstract  = {This paper introduces ENGLAWI, a large, versatile,
XML-encoded machine-readable dictionary extracted from Wiktionary.
ENGLAWI contains 752,769 articles encoding the full body of
information included in Wiktionary: simple words, compounds and
multiword expressions, lemmas and inflectional paradigms,
```

etymologies, phonemic transcriptions in IPA, definition glosses and usage examples, translations, semantic and morphological relations, spelling variants, etc. It is fully documented, released under a free license and supplied with G-PeTo, a series of scripts allowing easy information extraction from ENGLAWI. Additional resources extracted from ENGLAWI, such as an inflectional lexicon, a lexicon of diatopic variants and the inclusion dates of headwords in Wiktionary's nomenclature are also provided. The paper describes the content of the resource and illustrates how it can be - and has been - used in previous studies. We finally introduce an ongoing work that computes lexicographic word embeddings from ENGLAWI's definitions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.369}
}

@InProceedings{gomes-EtAl:2020:LREC,
  author   = {Gomes, Inês and Correia, Rui and Ribeiro, Jorge and Freitas, João},
  title   = {Effort Estimation in Named Entity Tagging Tasks},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {298--306},
  abstract = {Named Entity Recognition (NER) is an essential component of many Natural Language Processing pipelines. However, building these language dependent models requires large amounts of annotated data. Crowdsourcing emerged as a scalable solution to collect and enrich data in a more time-efficient manner. To manage these annotations at scale, it is important to predict completion timelines and compute fair pricing for workers in advance. To achieve these goals, we need to know how much effort will be taken to complete each task. In this paper, we investigate which variables influence the time spent on a named entity annotation task by a human. Our results are two-fold: first, the understanding of the effort-impacting factors which we divided into cognitive load and input length; and second, the performance of the prediction itself. On the latter, through model adaptation and feature engineering, we attained a Root Mean Squared Error (RMSE) of 25.68 words per minute with a Nearest Neighbors model.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.37}
}

@InProceedings{beniamine-maiden-round:2020:LREC,
  author   = {Beniamine, Sacha and Maiden, Martin and Round, Erich},
  title   = {Opening the Romance Verbal Inflection Dataset 2.0: A CLDF lexicon},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3027--3035},
  abstract  = {We introduce the Romance Verbal Inflection Dataset
2.0, a multilingual lexicon of Romance inflection covering 74
varieties. The lexicon provides verbal paradigm forms in broad IPA
phonemic notation. Both lexemes and paradigm cells are organized to
reflect cognacy. Such multi-lingual inflected lexicons annotated for
two dimensions of cognacy are necessary to study the evolution of
inflectional paradigms, and test linguistic hypotheses
systematically. However, these resources seldom exist, and when they
do, they are not usually encoded in computationally usable ways. The
Oxford Online Database of Romance Verb Morphology provides this kind
of information, however, it is not maintained anymore and is only
available as a web service without interfaces for machine-
readability. We collect its data and clean and correct it for
consistency using both heuristics and expert annotator judgements.
Most resources used to study language evolution computationally rely
strictly on multilingual contemporary information, and lack
information about prior stages of the languages. To provide such
information, we augment the database with Latin paradigms from the
LatInFlexi lexicon. Finally, to make it widely avalable, the
resource is released under a GPLv3 license in CLDF format.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.370}
}

@InProceedings{choe-park-kim:2020:LREC,
  author    = {Choe, Yo Joong  and  Park, Kyubyong  and  Kim,
Dongwoo},
  title     = {word2word: A Collection of Bilingual Lexicons for
3,564 Language Pairs},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3036--3045},
  abstract  = {We present word2word, a publicly available dataset
and an open-source Python package for cross-lingual word
translations extracted from sentence-level parallel corpora. Our
dataset provides top-k word translations in 3,564 (directed)
language pairs across 62 languages in OpenSubtitles2018 (Lison et
al., 2018). To obtain this dataset, we use a count-based bilingual
lexicon extraction model based on the observation that not only
source and target words but also source words themselves can be
highly correlated. We illustrate that the resulting bilingual
lexicons have high coverage and attain competitive translation
quality for several language pairs. We wrap our dataset and model in
an easy-to-use Python library, which supports downloading and
retrieving top-k word translations in any of the supported language
pairs as well as computing top-k word translations for custom
parallel corpora.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.371}
```

```
}

@InProceedings{cartoni-EtAl:2020:LREC,
  author    = {Cartoni, Bruno  and  Calvelo Aros, Daniel  and
Vrandecic, Denny  and  Lertpradit, Saran},
  title     = {Introducing Lexical Masks: a New Representation of
Lexical Entries for Better Evaluation and Exchange of Lexicons},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3046--3052},
  abstract  = {The evaluation and exchange of large lexicon
databases remains a challenge in many NLP applications. Despite the
existence of commonly accepted standards for the format and the
features used in a lexicon, there is still a lack of precise and
interoperable specification requirements about how lexical entries
of a particular language should look like, both in terms of the
numbers of forms and in terms of features associated with these
forms. This paper presents the notion of “lexical masks”, a powerful
tool used to evaluate and exchange lexicon databases in many
languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.372}
}

@InProceedings{alkhalil-habash-jiang:2020:LREC,
  author    = {Al Khalil, Muhamed  and  Habash, Nizar  and  Jiang,
Zhengyang},
  title     = {A Large-Scale Leveled Readability Lexicon for
Standard Arabic},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3053--3062},
  abstract  = {We present a large-scale 26,000-lemma leveled
readability lexicon for Modern Standard Arabic. The lexicon was
manually annotated in triplicate by language professionals from
three regions in the Arab world. The annotations show a high degree
of agreement; and major differences were limited to regional
variations. Comparing lemma readability levels with their
frequencies provided good insights in the benefits and pitfalls of
frequency-based readability approaches. The lexicon will be publicly
available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.373}
}

@InProceedings{stein:2020:LREC,
  author    = {Stein, Achim},
  title     = {Preserving Semantic Information from Old
```

Dictionaries: Linking Senses of the 'Altfranzösisches Wörterbuch' to WordNet},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3063--3068},
  abstract = {Historical dictionaries of the pre-digital period are important resources for the study of older languages. Taking the example of the 'Altfranzösisches Wörterbuch', an Old French dictionary published from 1925 onwards, this contribution shows how the printed dictionaries can be turned into a more easily accessible and more sustainable lexical database, even though a full-text retro-conversion is too costly. Over 57,000 German sense definitions were identified in uncorrected OCR output. For verbs and nouns, 34,000 senses of more than 20,000 lemmas were matched with GermaNet, a semantic network for German, and, in a second step, linked to synsets of the English WordNet. These results are relevant for the automatic processing of Old French, for the annotation and exploitation of Old French text corpora, and for the philological study of Old French in general.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.374}
}

@InProceedings{lai-winterstein:2020:LREC,
  author   = {Lai, Regine and Winterstein, Grégoire},
  title   = {Cifu: a Frequency Lexicon of Hong Kong Cantonese},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3069--3077},
  abstract = {This paper introduces Cifu, a lexical database for Hong Kong Cantonese (HKC) that offers phonological and orthographic information, frequency measures, and lexical neighborhood information for lexical items in HKC. Cifu is of use for NLP applications and the design and analysis of psycholinguistics experiments on HKC. We elaborate on the characteristics and challenges specific to HKC that were relevant in the design of Cifu. This includes lexical, orthographic and phonological aspects of HKC, word segmentation issues, the place of HKC in written media, and the availability of data. We discuss the measure of Neighborhood Density (ND), highlighting how the analytic nature of Cantonese and its writing system affect that measure. We justify using six different variations of ND, based on the possibility of inserting or deleting phonemes when searching for neighbors and on the choice of data for retrieving frequencies. Statistics about the four genres (written, adult spoken, children spoken and child-directed) within the dataset are discussed. We find that the lexical diversity of the child-directed speech genre is particularly low, compared to a size-

matched written corpus. The correlations of word frequencies of different genres are all high, but in generally decrease as word length increases.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.375}
}

@InProceedings{sprugnoli-EtAl:2020:LREC,
  author   = {Sprugnoli, Rachele and Passarotti, Marco and Corbetta, Daniela and Peverelli, Andrea},
  title   = {Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin.},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3078--3086},
  abstract = {Sentiment lexicons are essential for developing automatic sentiment analysis systems, but the resources currently available mostly cover modern languages. Lexicons for ancient languages are few and not evaluated with high-quality gold standards. However, the study of attitudes and emotions in ancient texts is a growing field of research which poses specific issues (e.g., lack of native speakers, limited amount of data, unusual textual genres for the sentiment analysis task, such as philosophical or documentary texts) and can have an impact on the work of scholars coming from several disciplines besides computational linguistics, e.g. historians and philologists. The work presented in this paper aims at providing the research community with a set of sentiment lexicons built by taking advantage of manually-curated resources belonging to the long tradition of Latin corpora and lexicons creation. Our interdisciplinary approach led us to release: i) two automatically generated sentiment lexicons; ii) a gold standard developed by two Latin language and culture experts; iii) a silver standard in which semantic and derivational relations are exploited so to extend the list of lexical items of the gold standard. In addition, the evaluation procedure is described together with a first application of the lexicons to a Latin tragedy.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.376}
}

@InProceedings{mohammad:2020:LREC2,
  author   = {Mohammad, Saif M.},
  title   = {WordWars: A Dataset to Examine the Natural Selection of Words},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3087--3095},

```
  abstract  = {There is a growing body of work on how word meaning
changes over time: mutation. In contrast, there is very little work
on how different words compete to represent the same meaning, and
how the degree of success of words in that competition changes over
time: natural selection. We present a new dataset, WordWars, with
historical frequency data from the early 1800s to the early 2000s
for monosemous English words in over 5000 synsets. We explore three
broad questions with the dataset: (1) what is the degree to which
predominant words in these synsets have changed, (2) how do
prominent word features such as frequency, length, and concreteness
impact natural selection, and (3) what are the differences between
the predominant words of the 2000s and the predominant words of
early 1800s. We show that close to one third of the synsets undergo
a change in the predominant word in this time period. Manual
annotation of these pairs shows that about 15\% of these are
orthographic variations, 25\% involve affix changes, and 60\% have
completely different roots. We find that frequency, length, and
concreteness all impact natural selection, albeit in different
ways.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.377}
}

@InProceedings{kanojia-EtAl:2020:LREC,
  author    = {Kanojia, Diptesh  and  Kulkarni, Malhar  and
Bhattacharyya, Pushpak  and  Haffari, Gholamreza},
  title     = {Challenge Dataset of Cognates and False Friend Pairs
from Indian Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3096--3102},
  abstract  = {Cognates are present in multiple variants of the same
text across different languages (e.g., "hund" in German and "hound"
in the English language mean "dog"). They pose a challenge to
various Natural Language Processing (NLP) applications such as
Machine Translation, Cross-lingual Sense Disambiguation,
Computational Phylogenetics, and Information Retrieval. A possible
solution to address this challenge is to identify cognates across
language pairs. In this paper, we describe the creation of two
cognate datasets for twelve Indian languages namely Sanskrit, Hindi,
Assamese, Oriya, Kannada, Gujarati, Tamil, Telugu, Punjabi, Bengali,
Marathi, and Malayalam. We digitize the cognate data from an Indian
language cognate dictionary and utilize linked Indian language
Wordnets to generate cognate sets. Additionally, we use the Wordnet
data to create a False Friends' dataset for eleven language pairs.
We also evaluate the efficacy of our dataset using previously
available baseline cognate detection approaches. We also perform a
manual evaluation with the help of lexicographers and release the
curated gold-standard dataset with this paper.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.378}
}
```

@InProceedings{iwai-EtAl:2020:LREC2,
  author   = {Iwai, Ritsuko and Kawahara, Daisuke and Kumada, Takatsune and Kurohashi, Sadao},
  title   = {Development of a Japanese Personality Dictionary based on Psychological Methods},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3103--3108},
  abstract = {We propose a new approach to constructing a personality dictionary with psychological evidence. In this study, we collect personality words, using word embeddings, and construct a personality dictionary with weights for Big Five traits. The weights are calculated based on the responses of the large sample (N=1,938, female = 1,004, M=49.8years old:20-78, SD=16.3). All the respondents answered a 20-item personality questionnaire and 537 personality items derived from word embeddings. We present the procedures to examine the qualities of responses with psychological methods and to calculate the weights. These result in a personality dictionary with two sub-dictionaries. We also discuss an application of the acquired resources.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.379}
}

@InProceedings{rodosthenous-EtAl:2020:LREC,
  author   = {Rodosthenous, Christos and Lyding, Verena and Sangati, Federico and König, Alexander and ul Hassan, Umair and Nicolas, Lionel and Horbacauskiene, Jolita and Katinskaia, Anisia and Aparaschivei, Lavinia},
  title   = {Using Crowdsourced Exercises for Vocabulary Training to Expand ConceptNet},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {307--316},
  abstract = {In this work, we report on a crowdsourcing experiment conducted using the V-TREL vocabulary trainer which is accessed via a Telegram chatbot interface to gather knowledge on word relations suitable for expanding ConceptNet. V-TREL is built on top of a generic architecture implementing the implicit crowdsourding paradigm in order to offer vocabulary training exercises generated from the commonsense knowledge-base ConceptNet and -- in the background -- to collect and evaluate the learners' answers to extend ConceptNet with new words. In the experiment about 90 university students learning English at C1 level, based on Common European Framework of Reference for Languages (CEFR), trained their vocabulary with V-TREL over a period of 16 calendar days. The

experiment allowed to gather more than 12,000 answers from learners on different question types. In this paper we present in detail the experimental setup and the outcome of the experiment, which indicates the potential of our approach for both crowdsourcing data as well as fostering vocabulary skills.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.38}
}

@InProceedings{islam-xiao-mercer:2020:LREC,
  author   = {Islam, Jumayel and Xiao, Lu and Mercer, Robert E.},
  title   = {A Lexicon-Based Approach for Detecting Hedges in Informal Text},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3109--3113},
  abstract = {Hedging is a commonly used strategy in conversational management to show the speaker's lack of commitment to what they communicate, which may signal problems between the speakers. Our project is interested in examining the presence of hedging words and phrases in identifying the tension between an interviewer and interviewee during a survivor interview. While there have been studies on hedging detection in the natural language processing literature, all existing work has focused on structured texts and formal communications. Our project thus investigated a corpus of eight unstructured conversational interviews about the Rwanda Genocide and identified hedging patterns in the interviewees' responses. Our work produced three manually constructed lists of hedge words, booster words, and hedging phrases. Leveraging these lexicons, we developed a rule-based algorithm that detects sentence-level hedges in informal conversations such as survivor interviews. Our work also produced a dataset of 3000 sentences having the categories Hedge and Non-hedge annotated by three researchers. With experiments on this annotated dataset, we verify the efficacy of our proposed algorithm. Our work contributes to the further development of tools that identify hedges from informal conversations and discussions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.380}
}

@InProceedings{nishihara-kajiwara:2020:LREC,
  author   = {Nishihara, Daiki and Kajiwara, Tomoyuki},
  title   = {Word Complexity Estimation for Japanese Lexical Simplification},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {3114--3120},
  abstract  = {We introduce three language resources for Japanese
lexical simplification: 1) a large-scale word complexity lexicon, 2)
the first synonym lexicon for converting complex words to simpler
ones, and 3) the first toolkit for developing and benchmarking
Japanese lexical simplification system. Our word complexity lexicon
is expanded to a broader vocabulary using a classifier trained on a
small, high-quality word complexity lexicon created by Japanese
language teachers. Based on this word complexity estimator, we
extracted simplified word pairs from a large-scale synonym lexicon
and constructed a simplified synonym lexicon useful for lexical
simplification. In addition, we developed a Python library that
implements automatic evaluation and key methods in each subtask to
ease the construction of a lexical simplification pipeline.
Experimental results show that the proposed method based on our
lexicon achieves the highest performance of Japanese lexical
simplification. The current lexical simplification is mainly studied
in English, which is rich in language resources such as lexicons and
toolkits. The language resources constructed in this study will help
advance the lexical simplification system in Japanese.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.381}
}

@InProceedings{huo-demelo:2020:LREC,
  author    = {Huo, Da  and  de Melo, Gerard},
  title     = {Inducing Universal Semantic Tag Vectors},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3121--3127},
  abstract  = {Given the well-established usefulness of part-of-
speech tag annotations in many syntactically oriented downstream NLP
tasks, the recently proposed notion of semantic tagging (Bjerva et
al. 2016) aims at tagging words with tags informed by semantic
distinctions, which are likely to be useful across a range of
semantic tasks. To this end, their annotation scheme distinguishes,
for instance, privative attributes from subsective ones. While
annotated corpora exist, their size is limited and thus many words
are out-of-vocabulary words. In this paper, we study to what extent
we can automatically predict the tags associated with unseen words.
We draw on large-scale word representation data to derive a large
new Semantic Tag lexicon. Our experiments show that we can infer
semantic tags for words with high accuracy both monolingually and
cross-lingually.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.382}
}

@InProceedings{coole-rayson-mariani:2020:LREC,
  author    = {Coole, Matthew  and  Rayson, Paul  and  Mariani,
John},
  title     = {LexiDB: Patterns \& Methods for Corpus Linguistic
```

Database Management},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3128--3135},
  abstract = {LexiDB is a tool for storing, managing and querying corpus data. In contrast to other database management systems (DBMSs), it is designed specifically for text corpora. It improves on other corpus management systems (CMSs) because data can be added and deleted from corpora on the fly with the ability to add live data to existing corpora. LexiDB sits between these two categories of DBMSs and CMSs, more specialised to language data than a general purpose DBMS but more flexible than a traditional static corpus management system. Previous work has demonstrated the scalability of LexiDB in response to the growing need to be able to scale out for ever growing corpus datasets. Here, we present the patterns and methods developed in LexiDB for storage, retrieval and querying of multi-level annotated corpus data. These techniques are evaluated and compared to an existing CMS (Corpus Workbench CWB - CQP) and indexer (Lucene). We find that LexiDB consistently outperforms existing tools for corpus queries. This is particularly apparent with large corpora and when handling queries with large result sets},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.383}
}

@InProceedings{kettnerov-EtAl:2020:LREC,
  author   = {Kettnerová, Václava and Lopatkova, Marketa and Vernerová, Anna and Barancikova, Petra},
  title   = {Towards a Semi-Automatic Detection of Reflexive and Reciprocal Constructions and Their Representation in a Valency Lexicon},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3136--3144},
  abstract = {Valency lexicons usually describe valency behavior of verbs in non-reflexive and non-reciprocal constructions. However, reflexive and reciprocal constructions are common morphosyntactic forms of verbs. Both of these constructions are characterized by regular changes in morphosyntactic properties of verbs, thus they can be described by grammatical rules. On the other hand, the possibility to create reflexive and/or reciprocal constructions cannot be trivially derived from the morphosyntactic structure of verbs as it is conditioned by their semantic properties as well. A large-coverage valency lexicon allowing for rule based generation of all well formed verb constructions should thus integrate the information on reflexivity and reciprocity. In this paper, we

```
propose a semi-automatic procedure, based on grammatical constraints
on reflexivity and reciprocity, detecting those verbs that form
reflexive and reciprocal constructions in corpus data. However,
exploitation of corpus data for this purpose is complicated due to
the diverse functions of reflexive markers crossing the domain of
reflexivity and reciprocity. The list of verbs identified by the
previous procedure is thus further used in an automatic experiment,
applying word embeddings for detecting semantically similar verbs.
These candidate verbs have been manually verified and annotation of
their reflexive and reciprocal constructions has been integrated
into the valency lexicon of Czech verbs VALLEX.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.384}
}

@InProceedings{vidalgorne-decoursperez:2020:LREC,
  author    = {Vidal-Gorène, Chahan  and  Decours-Perez, Aliénor},
  title     = {Languages Resources for Poorly Endowed Languages :
The Case Study of Classical Armenian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3145--3152},
  abstract  = {Classical Armenian is a poorly endowed language, that
despite a great tradition of lexicographical erudition is coping
with a lack of resources. Although numerous initiatives exist to
preserve the Classical Armenian language, the lack of precise and
complete grammatical and lexicographical resources remains. This
article offers a situation analysis of the existing resources for
Classical Armenian and presents the new digital resources provided
on the Calfa platform. The Calfa project gathers existing resources
and updates, enriches and enhances their content to offer the
richest database for Classical Armenian today. Faced with the
challenges specific to a poorly endowed language, the Calfa project
is also developing new technologies and solutions to enable
preservation, advanced research, and larger systems and developments
for the Armenian language},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.385}
}

@InProceedings{takeuchi-EtAl:2020:LREC,
  author    = {Takeuchi, Koichi  and  Butler, Alastair  and
Nagasaki, Iku  and  Okamura, Takuya  and  Pardeshi, Prashant},
  title     = {Constructing Web-Accessible Semantic Role Labels and
Frames for Japanese as Additions to the NPCMJ Parsed Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3153--3161},
```

```
  abstract  = {As part of constructing the NINJAL Parsed Corpus of
Modern Japanese (NPCMJ), a web-accessible language resource, we are
adding frame information for predicates, together with two types of
semantic role labels that mark the contributions of arguments. One
role type consists of numbered semantic roles, like in PropBank, to
capture relations between arguments in different syntactic patterns.
The other role type consists of semantic roles with conventional
names.Both role types are compatible with hierarchical frames that
belong to related predicates. Adding semantic role and frame
information to the NPCMJ will support a web environment where
language learners and linguists can search examples of Japanese for
syntactic and semantic features. The annotation will also provide a
language resource for NLP researchers making semantic parsing models
(e.g., for AMR parsing) following machine learning approaches. In
this paper, we describe how the two types of semantic role labels
are defined under the frame based approach, i.e., both types can be
consistently applied when linked to corresponding frames. Then we
show special cases of syntactic patterns and the current status of
the annotation work.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.386}
}

@InProceedings{vossen-EtAl:2020:LREC,
  author    = {Vossen, Piek  and  Ilievski, Filip  and  Postma,
Marten  and  Fokkens, Antske  and  Minnema, Gosse  and  Remijnse,
Levi},
  title     = {Large-scale Cross-lingual Language Resources for
Referencing and Framing},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3162--3171},
  abstract  = {In this article, we lay out the basic ideas and
principles of the project Framing Situations in the Dutch Language.
We provide our first results of data acquisition, together with the
first data release. We introduce the notion of cross-lingual
referential corpora. These corpora consist of texts that make
reference to exactly the same incidents. The referential grounding
allows us to analyze the framing of these incidents in different
languages and across different texts. During the project, we will
use the automatically generated data to study linguistic framing as
a phenomenon, build framing resources such as lexicons and corpora.
We expect to capture larger variation in framing compared to
traditional approaches for building such resources. Our first data
release, which contains structured data about a large number of
incidents and reference texts, can be found at http://
dutchframenet.nl/data-releases/.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.387}
}

@InProceedings{khan-EtAl:2020:LREC,
```

```
  author    = {Khan, Fahad  and  Romary, Laurent  and  Salgado, Ana
and  Bowers, Jack  and  Khemakhem, Mohamed  and  Tasovac, Toma},
  title     = {Modelling Etymology in LMF/TEI: The Grande Dicionário
Houaiss da Língua Portuguesa Dictionary as a Use Case},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3172--3180},
  abstract  = {In this article we will introduce two of the new
parts of the new multi-part version of the Lexical Markup Framework
(LMF) ISO standard, namely part 3 of the standard (ISO 24613-3),
which deals with etymological and diachronic data, and Part 4 (ISO
24613-4), which consists of a TEI serialisation of all of the prior
parts of the model. We will demonstrate the use of both standards by
describing the LMF encoding of a small number of examples taken from
a sample conversion of the reference Portuguese dictionary
\textit{Grande Dicionário Houaiss da Língua Portuguesa}, part of a
broader experiment comprising the analysis of different,
heterogeneously encoded, Portuguese lexical resources. We present
the examples in the Unified Modelling Language (UML) and also in a
couple of cases in TEI.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.388}
}

@InProceedings{bond-EtAl:2020:LREC1,
  author    = {Bond, Francis  and  Nomoto, Hiroki  and  Morgado da
Costa, Luís  and  Bond, Arthur},
  title     = {Linking the TUFS Basic Vocabulary to the Open
Multilingual Wordnet},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3181--3188},
  abstract  = {We describe the linking of the TUFS Basic Vocabulary
Modules, created for online language learning, with the Open
Multilingual Wordnet. The TUFS modules have roughly 500 lexical
entries in 30 languages, each with the lemma, a link across the
languages, an example sentence, usage notes and sound files. The
Open Multilingual Wordnet has 34 languages (11 shared with TUFS)
organized into synsets linked by semantic relations, with examples
and definitions for some languages. The links can be used to (i)
evaluate existing wordnets, (ii) add data to these wordnets and
(iii) create new open wordnets for Khmer, Korean, Lao, Mongolian,
Russian, Tagalog, Urdua nd Vietnamese},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.389}
}

@InProceedings{bailly-EtAl:2020:LREC,
```

```
  author    = {Bailly, Gérard  and  Godde, Erika  and  Piat-
Marchand, Anne-Laure  and  Bosse, Marie-Line},
  title     = {Predicting Multidimensional Subjective Ratings of
Children' Readings from the Speech Signals for the Automatic
Assessment of Fluency},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {317--322},
  abstract  = {The objective of this research is to estimate
multidimensional subjective ratings of the reading performance of
young readers from signal-based objective measures. We here combine
linguistic features (number of correct words, repetitions,
deletions, insertions uttered per minute . . . ) with phonetic
features. Expressivity is particularly difficult to predict since
there is no unique golden standard. We here propose a novel
framework for performing such an estimation that exploits multiple
references performed by adults and demonstrate its efficiency using
recordings of 273 pupils.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.39}
}

@InProceedings{bond-EtAl:2020:LREC2,
  author    = {Bond, Francis  and  Morgado da Costa, Luis  and
Goodman, Michael Wayne  and  McCrae, John Philip  and  Lohk, Ahti},
  title     = {Some Issues with Building a Multilingual Wordnet},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3189--3197},
  abstract  = {In this paper we discuss the experience of bringing
together over 40 different wordnets. We introduce some extensions to
the GWA wordnet LMF format proposed in Vossen et al. (2016) and look
at how this new information can be displayed. Notable extensions
include: confidence, corpus frequency, orthographic variants,
lexicalized and non-lexicalized synsets and lemmas, new parts of
speech, and more. Many of these extensions already exist in multiple
wordnets — the challenge was to find a compatible representation. To
this end, we introduce a new version of the Open Multilingual
Wordnet (Bond and Foster, 2013), that integrates a new set of tools
that tests the extensions introduced by this new format, while also
ensuring the integrity of the Collaborative Interlingual Index
(CILI: Bond et al., 2016), avoiding the same new concept to be
introduced through multiple projects.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.390}
}

@InProceedings{khokhlova:2020:LREC,
```

```
  author    = {Khokhlova, Maria},
  title     = {Collocations in Russian Lexicography and Russian
Collocations Database},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3198--3206},
  abstract  = {The paper presents the issue of collocability and
collocations in Russian and gives a survey of a wide range of
dictionaries both printed and online ones that describe
collocations. Our project deals with building a database that will
include dictionary and statistical collocations. The former can be
described in various lexicographic resources whereas the latter can
be extracted automatically from corpora. Dictionaries differ among
themselves, the information is given in various ways, making it hard
for language learners and researchers to acquire data. A number of
dictionaries were analyzed and processed to retrieve verified
collocations, however the overlap between the lists of collocations
extracted from them is still rather small. This fact indicates there
is a need to create a unified resource which takes into account
collocability and more examples. The proposed resource will also be
useful for linguists and for studying Russian as a foreign language.
The obtained results can be important for machine learning and for
other NLP tasks, for instance, automatic clustering of word
combinations and disambiguation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.391}
}

@InProceedings{fourrier-sagot:2020:LREC,
  author    = {Fourrier, Clémentine  and  Sagot, Benoît},
  title     = {Methodological Aspects of Developing and Managing an
Etymological Lexical Resource: Introducing EtymDB-2.0},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3207--3216},
  abstract  = {Diachronic lexical information is not only important
in the field of historical linguistics, but is also increasingly
used in NLP, most recently for machine translation of low resource
languages. Therefore, there is a need for fine-grained, large-
coverage and accurate etymological lexical resources. In this paper,
we propose a set of guidelines to generate such resources, for each
step of the life-cycle of an etymological lexicon: creation, update,
evaluation, dissemination, and exploitation. To illustrate the
guidelines, we introduce EtymDB 2.0, an etymological database
automatically generated from the Wiktionary, which contains 1.8
million lexemes, linked by more than 700,000 fine-grained
etymological relations, across 2,536 living and dead languages. We
```

also introduce use cases for which EtymDB 2.0 could represent a key resource, such as phylogenetic tree generation, low resource machine translation or medieval languages study.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.392}
}

@InProceedings{guibon-sagot:2020:LREC,
  author   = {Guibon, Gaël and Sagot, Benoît},
  title   = {OFrLex: A Computational Morphological and Syntactic Lexicon for Old French},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3217--3225},
  abstract = {In this paper we describe our work on the development and enrichment of OFrLex, a freely available, large-coverage morphological and syntactic Old French lexicon. We rely on several heterogeneous language resources to extract structured and exploitable information. The extraction follows a semi-automatic procedure with substantial manual steps to respond to difficulties encountered while aligning lexical entries from distinct language resources. OFrLex aims at improving natural language processing tasks on Old French such as part-of-speech tagging and dependency parsing. We provide quantitative information on OFrLex and discuss its reliability. We also describe and evaluate a semi-automatic, word-embedding-based lexical enrichment process aimed at increasing the accuracy of the resource. Results of this extension technique will be manually validated in the near future, a step that will take advantage of OFrLex's viewing, searching and editing interface, which is already accessible online.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.393}
}

@InProceedings{ciobanu-dinu-zoicas:2020:LREC,
  author   = {Ciobanu, Alina Maria and Dinu, Liviu P. and Zoicas, Laurentiu},
  title   = {Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3226--3231},
  abstract = {Producing related words is a key concern in historical linguistics. Given an input word, the task is to automatically produce either its proto-word, a cognate pair or a modern word derived from it. In this paper, we apply a method for producing related words based on sequence labeling, aiming to fill in the gaps in incomplete cognate sets in Romance languages with

```
Latin etymology (producing Romanian cognates that are missing) and
to reconstruct uncertified Latin words. We further investigate an
ensemble-based aggregation for combining and re-ranking the word
productions of multiple languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.394}
}

@InProceedings{ahmadi-EtAl:2020:LREC,
  author    = {Ahmadi, Sina  and  McCrae, John Philip  and  Nimb,
Sanni  and  Khan, Fahad  and  Monachini, Monica  and  Pedersen,
Bolette  and  Declerck, Thierry  and  Wissik, Tanja  and  Bellandi,
Andrea  and  Pisani, Irene  and  Troelsgård, Thomas  and  Olsen,
Sussi  and  Krek, Simon  and  Lipp, Veronika  and  Váradi, Tamás
and  Simon, László  and  Gyorffy, András  and  Tiberius, Carole  and
Schoonheim, Tanneke  and  Ben Moshe, Yifat  and  Rudich, Maya  and
Abu Ahmad, Raya  and  Lonke, Dorielle  and  Kovalenko, Kira  and
Langemets, Margit  and  Kallas, Jelena  and  Dereza, Oksana  and
Fransen, Theodorus  and  Cillessen, David  and  Lindemann, David
and  Alonso, Mikel  and  Salgado, Ana  and  Luis Sancho, José  and
Ureña-Ruiz, Rafael-J.  and  Porta Zamorano, Jordi  and  Simov, Kiril
and  Osenova, Petya  and  Kancheva, Zara  and  Radev, Ivaylo  and
Stanković, Ranka  and  Perdih, Andrej  and  Gabrovsek, Dejan},
  title     = {A Multilingual Evaluation Dataset for Monolingual
Word Sense Alignment},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3232--3242},
  abstract  = {Aligning senses across resources and languages is a
challenging task with beneficial applications in the field of
natural language processing and electronic lexicography. In this
paper, we describe our efforts in manually aligning monolingual
dictionaries. The alignment is carried out at sense-level for
various resources in 15 languages. Moreover, senses are annotated
with possible semantic relationships such as broadness, narrowness,
relatedness, and equivalence. In comparison to previous datasets for
this task, this dataset covers a wide range of languages and
resources and focuses on the more challenging task of linking
general-purpose language. We believe that our data will pave the way
for further advances in alignment and evaluation of word senses by
creating new solutions, particularly those notoriously requiring
data such as neural networks. Our resources are publicly available
at https://github.com/elexis-eu/MWSA.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.395}
}

@InProceedings{allen-EtAl:2020:LREC,
  author    = {Allen, James  and  An, Hannah  and  Bose, Ritwik  and
de Beaumont, Will  and  Teng, Choh Man},
  title     = {A Broad-Coverage Deep Semantic Lexicon for Verbs},
  booktitle      = {Proceedings of The 12th Language Resources and
```

```
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3243--3251},
  abstract  = {Progress on deep language understanding is inhibited
by the lack of a broad coverage lexicon that connects linguistic
behavior to ontological concepts and axioms. We have developed
COLLIE-V, a deep lexical resource for verbs, with the coverage of
WordNet and syntactic and semantic details that meet or exceed
existing resources. Bootstrapping from a hand-built lexicon and
ontology, new ontological concepts and lexical entries, together
with semantic role preferences and entailment axioms, are
automatically derived by combining multiple constraints from parsing
dictionary definitions and examples. We evaluated the accuracy of
the technique along a number of different dimensions and were able
to obtain high accuracy in deriving new concepts and lexical
entries. COLLIE-V is publicly available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.396}
}

@InProceedings{wu-yarowsky:2020:LREC,
  author    = {Wu, Winston  and  Yarowsky, David},
  title     = {Computational Etymology and Word Emergence},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3252--3259},
  abstract  = {We developed an extensible, comprehensive Wiktionary
parser that improves over several existing parsers. We predict the
etymology of a word across the full range of etymology types and
languages in Wiktionary, showing improvements over a strong
baseline. We also model word emergence and show the application of
etymology in modeling this phenomenon. We release our parser to
further research in this understudied field.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.397}
}

@InProceedings{rudnicka-naskrt:2020:LREC,
  author    = {Rudnicka, Ewa  and  Naskręt, Tomasz},
  title     = {A Dataset of Translational Equivalents Built on the
Basis of plWordNet-Princeton WordNet Synset Mapping},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3260--3264},
  abstract  = {The paper presents a dataset of 11,000 Polish-English
```

translational equivalents in the form of pairs of plWordNet and Princeton WordNet lexical units linked by three types of equivalence links: strong equivalence, regular equivalence, and weak equivalence. The resource consists of the two subsets. The first subset was built in result of manual annotation of an extended sample of Polish-English sense pairs partly randomly extracted from synsets linked by interlingual relations such as I-synononymy, I-partial synonymy and I-hyponymy and partly manually selected from the surrounding synsets in the hypernymy hierarchy. The second subset was created as a result of the manual checkup of an automatically generated lists of pairs of sense equivalents on the basis of a couple of simple, rule-based heuristics. For both subsets, the same methodology of equivalence annotation was adopted based on the verification of a set of formal, semantic-pragmatic and translational features. The constructed dataset is a novum in the wordnet domain and can facilitate the precision of bilingual NLP tasks such as automatic translation, bilingual word sense disambiguation and sentiment annotation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.398}
}

@InProceedings{benites-EtAl:2020:LREC,
  author   = {Benites, Fernando and Duivesteijn, Gilbert François and von Däniken, Pius and Cieliebak, Mark},
  title   = {TRANSLIT: A Large-scale Name Transliteration Resource},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3265--3271},
  abstract = {Transliteration is the process of expressing a proper name from a source language in the characters of a target language (e.g. from Cyrillic to Latin characters). We present TRANSLIT, a large-scale corpus with approx. 1.6 million entries in more than 180 languages with about 3 million variations of person and geolocation names. The corpus is based on various public data sources, which have been transformed into a unified format to simplify their usage, plus a newly compiled dataset from Wikipedia. In addition, we apply several machine learning methods to establish baselines for automatically detecting transliterated names in various languages. Our best systems achieve an accuracy of 92\% on identification of transliterated pairs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.399}
}

@InProceedings{henlein-mehler:2020:LREC,
  author   = {Henlein, Alexander and Mehler, Alexander},
  title   = {On the Influence of Coreference Resolution on Word Embeddings in Lexical-semantic Evaluation Tasks},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {27--33},
  abstract  = {Coreference resolution (CR) aims to find all spans of
a text that refer to the same entity. The F1-Scores on these task
have been greatly improved by new developed End2End-approaches and
transformer networks. The inclusion of CR as a pre-processing step
is expected to lead to improvements in downstream tasks. The paper
examines this effect with respect to word embeddings. That is, we
analyze the effects of CR on six different embedding methods and
evaluate them in the context of seven lexical-semantic evaluation
tasks and instantiation/hypernymy detection. Especially in the last
tasks we hoped for a significant increase in performance. We show
that all word embedding approaches do not benefit significantly from
pronoun substitution. The measurable improvements are only marginal
(around 0.5\% in most test cases). We explain this result with the
loss of contextual information, reduction of the relative occurrence
of rare words and the lack of pronouns to be replaced.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.4}
}

@InProceedings{akhlaghi-EtAl:2020:LREC,
  author    = {Akhlaghi, Elham  and  Bédi, Branislav  and  Bektaş,
Fatih  and  Berthelsen, Harald  and  Butterweck, Matthias  and
Chua, Cathy  and  Cucchiarin, Catia  and  Eryiğit, Gülşen  and
Gerlach, Johanna  and  Habibi, Hanieh  and  Ní Chiaráin, Neasa  and
Rayner, Manny  and  Steingrímsson, Steinþór  and  Strik, Helmer},
  title     = {Constructing Multimodal Language Learner Texts Using
LARA: Experiences with Nine Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {323--331},
  abstract  = {LARA (Learning and Reading Assistant) is an open
source platform whose purpose is to support easy conversion of plain
texts into multimodal online versions suitable for use by language
learners. This involves semi-automatically tagging the text, adding
other annotations and recording audio. The platform is suitable for
creating texts in multiple languages via crowdsourcing techniques
that can be used for teaching a language via reading and listening.
We present results of initial experiments by various collaborators
where we measure the time required to produce substantial LARA
resources, up to the length of short novels, in Dutch, English,
Farsi, French, German, Icelandic, Irish, Swedish and Turkish. The
first results are encouraging. Although there are some startup
problems, the conversion task seems manageable for the languages
tested so far. The resulting enriched texts are posted online and
are freely available in both source and compiled form.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.40}
```

}

@InProceedings{lmjeronimo-EtAl:2020:LREC,
  author   = {L. M. Jeronimo, Caio and E. C. Campelo, Claudio and Balby Marinho, Leandro and Sales, Allan and Veloso, Adriano and Viola, Roberta},
  title   = {Computing with Subjectivity Lexicons},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3272--3280},
  abstract = {In this paper, we introduce a new set of lexicons for expressing subjectivity in text documents written in Brazilian Portuguese. Besides the non-English idiom, in contrast to other subjectivity lexicons available, these lexicons represent different subjectivity dimensions (other than sentiment) and are more compact in number of terms. This last feature was designed intentionally to leverage the power of word embedding techniques, i.e., with the words mapped to an embedding space and the appropriate distance measures, we can easily capture semantically related words to the ones in the lexicons. Thus, we do not need to build comprehensive vocabularies and can focus on the most representative words for each lexicon dimension. We showcase the use of these lexicons in three highly non-trivial tasks: (1) Automated Essay Scoring in the Presence of Biased Ratings, (2) Subjectivity Bias in Brazilian Presidential Elections and (3) Fake News Classification Based on Text Subjectivity. All these tasks involve text documents written in Portuguese.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.400}
}

@InProceedings{chiarcos-fth-ionov:2020:LREC,
  author   = {Chiarcos, Christian and Fäth, Christian and Ionov, Maxim},
  title   = {The ACoLi Dictionary Graph},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3281--3290},
  abstract = {In this paper, we report the release of the ACoLi Dictionary Graph, a large-scale collection of multilingual open source dictionaries available in two machine-readable formats, a graph representation in RDF, using the OntoLex-Lemon vocabulary, and a simple tabular data format to facilitate their use in NLP tasks, such as translation inference across dictionaries. We describe the mapping and harmonization of the underlying data structures into a unified representation, its serialization in RDF and TSV, and the release of a massive and coherent amount of lexical data under open

```
licenses.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.401}
}

@InProceedings{midriganciochina-EtAl:2020:LREC,
  author    = {Midrigan - Ciochina, Ludmila  and  Boyd, Victoria
and  Sanchez-Ortega, Lucila  and  Malancea\_Malac, Diana  and
Midrigan, Doina  and  Corina, David P.},
  title     = {Resources in Underrepresented Languages: Building a
Representative Romanian Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3291--3296},
  abstract  = {The effort in the field of Linguistics to develop
theories that aim to explain language-dependent effects on language
processing is greatly facilitated by the availability of reliable
resources representing different languages. This project presents a
detailed description of the process of creating a large and
representative corpus in Romanian - a relatively under-resourced
language with unique structural and typological characteristics,
that can be used as a reliable language resource for linguistic
studies. The decisions that have guided the construction of the
corpus, including the type of corpus, its size and component
resource files are discussed. Issues related to data collection,
data organization and storage, as well as characteristics of the
data included in the corpus are described. Currently, the corpus has
approximately 5,500,000 tokens originating from written text and
100,000 tokens of spoken language. it includes language samples that
represent a wide variety of registers (i.e. written language - 16
registers and 5 registers of spoken language), as well as different
authors and speakers},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.402}
}

@InProceedings{kirchmeier-EtAl:2020:LREC,
  author    = {Kirchmeier, Sabine  and  Pedersen, Bolette  and
Nimb, Sanni  and  Diderichsen, Philip  and  Henrichsen, Peter Juel},
  title     = {World Class Language Technology - Developing a
Language Technology Strategy for Danish},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3297--3301},
  abstract  = {Although Denmark is one of the most digitized
countries in Europe, no coordinated efforts have been made in recent
years to support the Danish language with regard to language
technology and artificial intelligence. In March 2019, however, the
```

```
Danish government adopted a new, ambitious strategy for LT and
artificial intelligence. In this paper, we describe the process
behind the development of the language-related parts of the
strategy: A Danish Language Technology Committee was constituted and
a comprehensive series of workshops were organized in which users,
suppliers, developers, and researchers gave their valuable input
based on their experiences. We describe how, based on this
experience, the focus areas and recommendations for the LT strategy
were established, and which steps are currently taken in order to
put the strategy into practice.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.403}
}

@InProceedings{battisti-EtAl:2020:LREC,
  author    = {Battisti, Alessia  and  Pfütze, Dominik  and
Säuberli, Andreas  and  Kostrzewa, Marek  and  Ebling, Sarah},
  title     = {A Corpus for Automatic Readability Assessment and
Text Simplification of German},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3302--3311},
  abstract  = {In this paper, we present a corpus for use in
automatic readability assessment and automatic text simplification
for German, the first of its kind for this language. The corpus is
compiled from web sources and consists of parallel as well as
monolingual-only (simplified German) data amounting to approximately
6,200 documents (nearly 211,000 sentences). As a unique feature, the
corpus contains information on text structure (e.g., paragraphs,
lines), typography (e.g., font type, font style), and images
(content, position, and dimensions). While the importance of
considering such information in machine learning tasks involving
simplified language, such as readability assessment, has repeatedly
been stressed in the literature, we provide empirical evidence for
its benefit. We also demonstrate the added value of leveraging
monolingual-only data for automatic text simplification via machine
translation through applying back-translation, a data augmentation
technique.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.404}
}

@InProceedings{vandenheuvel-EtAl:2020:LREC1,
  author    = {van den Heuvel, Henk  and  Oostdijk, Nelleke  and
Rowland, Caroline  and  Trilsbeek, Paul},
  title     = {The CLARIN Knowledge Centre for Atypical
Communication Expertise},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {3312--3316},
  abstract  = {This paper introduces a new CLARIN Knowledge Center
which is the K-Centre for Atypical Communication Expertise (ACE for
short) which has been established at the Centre for Language and
Speech Technology (CLST) at Radboud University. Atypical
communication is an umbrella term used here to denote language use
by second language learners, people with language disorders or those
suffering from language disabilities, but also more broadly by
bilinguals and users of sign languages. It involves multiple
modalities (text, speech, sign, gesture) and encompasses different
developmental stages. ACE closely collaborates with The Language
Archive (TLA) at the Max Planck Institute for Psycholinguistics in
order to safeguard GDPR-compliant data storage and access. We
explain the mission of ACE and show its potential on a number of
showcases and a use case.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.405}
}

@InProceedings{vandenheuvel-EtAl:2020:LREC2,
  author    = {van den Heuvel, Henk  and  Kelli, Aleksei  and
Klessa, Katarzyna  and  Salaasti, Satu},
  title     = {Corpora of Disordered Speech in the Light of the
GDPR: Two Use Cases from the DELAD Initiative},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3317--3321},
  abstract  = {Corpora of disordered speech (CDS) are costly to
collect and difficult to share due to personal data protection and
intellectual property (IP) issues. In this contribution we discuss
the legal grounds for processing CDS in the light of the GDPR, and
illustrate these with two use cases from the DELAD context. One use
case deals with clinical datasets and another with legacy data from
Polish hearing-impaired children. For both cases, processing based
on consent and on public interest are taken into consideration.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.406}
}

@InProceedings{rehm-EtAl:2020:LREC1,
  author    = {Rehm, Georg  and  Marheinecke, Katrin  and  Hegele,
Stefanie  and  Piperidis, Stelios  and  Bontcheva, Kalina  and
Hajic, Jan  and  Choukri, Khalid  and  Vasiļjevs, Andrejs  and
Backfried, Gerhard  and  Prinz, Christoph  and  Gomez-Perez, Jose
Manuel  and  Meertens, Luc  and  Lukowicz, Paul  and  van Genabith,
Josef  and  Lösch, Andrea  and  Slusallek, Philipp  and  Irgens,
Morten  and  Gatellier, Patrick  and  Köhler, Joachim  and  Le Bars,
Laure  and  Anastasiou, Dimitra  and  Auksoriūtė, Albina  and  Bel,
Núria  and  Branco, António  and  Budin, Gerhard  and  Daelemans,
Walter  and  De Smedt, Koenraad  and  Garabík, Radovan  and
Gavriilidou, Maria  and  Gromann, Dagmar  and  Koeva, Svetla  and
```

Krek, Simon and Krstev, Cvetana and Lindén, Krister and Magnini, Bernardo and Odijk, Jan and Ogrodniczuk, Maciej and Rögnvaldsson, Eiríkur and Rosner, Mike and Pedersen, Bolette and Skadina, Inguna and Tadić, Marko and Tufiș, Dan and Váradi, Tamás and Vider, Kadri and Way, Andy and Yvon, François},
  title   = {The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3322--3332},
  abstract = {Multilingualism is a cultural cornerstone of Europe and firmly anchored in the European treaties including full language equality. However, language barriers impacting business, cross-lingual and cross-cultural communication are still omnipresent. Language Technologies (LTs) are a powerful means to break down these barriers. While the last decade has seen various initiatives that created a multitude of approaches and technologies tailored to Europe’s specific needs, there is still an immense level of fragmentation. At the same time, AI has become an increasingly important concept in the European Information and Communication Technology area. For a few years now, AI – including many opportunities, synergies but also misconceptions – has been overshadowing every other topic. We present an overview of the European LT landscape, describing funding programmes, activities, actions and challenges in the different countries with regard to LT, including the current state of play in industry and the LT market. We present a brief overview of the main LT-related activities on the EU level in the last ten years and develop strategic guidance with regard to four key dimensions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.407}
}

@InProceedings{gilliswebber-tittel:2020:LREC,
  author   = {Gillis-Webber, Frances and Tittel, Sabine},
  title   = {A Framework for Shared Agreement of Language Tags beyond ISO 639},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3333--3339},
  abstract = {The identification and annotation of languages in an unambiguous and standardized way is essential for the description of linguistic data. It is the prerequisite for machine-based interpretation, aggregation, and re-use of the data with respect to different languages. This makes it a key aspect especially for

Linked Data and the multilingual Semantic Web. The standard for language tags is defined by IETF’s BCP 47 and ISO 639 provides the language codes that are the tags’ main constituents. However, for the identification of lesser-known languages, endangered languages, regional varieties or historical stages of a language, the ISO 639 codes are insufficient. Also, the optional language sub-tags compliant with BCP 47 do not offer a possibility fine-grained enough to represent linguistic variation. We propose a versatile pattern that extends the BCP 47 sub-tag 'privateuse' and is, thus, able to overcome the limits of BCP 47 and ISO 639. Sufficient coverage of the pattern is demonstrated with the use case of linguistic Linked Data of the endangered Gascon language. We show how to use a URI shortcode for the extended sub-tag, making the length compliant with BCP 47. We achieve this with a web application and API developed to encode and decode the language tag.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.408}
}

@InProceedings{krek-EtAl:2020:LREC,
  author   = {Krek, Simon and Arhar Holdt, Špela and Erjavec, Tomaž and Čibej, Jaka and Repar, Andraz and Gantar, Polona and Ljubešić, Nikola and Kosem, Iztok and Dobrovoljc, Kaja},
  title   = {Gigafida 2.0: The Reference Corpus of Written Standard Slovene},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3340--3345},
  abstract = {We describe a new version of the Gigafida reference corpus of Slovene. In addition to updating the corpus with new material and annotating it with better tools, the focus of the upgrade was also on its transformation from a general reference corpus, which contains all language variants including non-standard language, to the corpus of standard (written) Slovene. This decision could be implemented as new corpora dedicated specifically to non-standard language emerged recently. In the new version, the whole Gigafida corpus was deduplicated for the first time, which facilitates automatic extraction of data for the purposes of compilation of new lexicographic resources such as the collocations dictionary and the thesaurus of Slovene.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.409}
}

@InProceedings{pilan-EtAl:2020:LREC,
  author   = {Pilan, Ildiko and Lee, John and Yeung, Chak Yan and Webster, Jonathan},
  title   = {A Dataset for Investigating the Impact of Feedback on Student Revision Outcome},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},

```
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {332--339},
  abstract  = {We present an annotation scheme and a dataset of
teacher feedback provided for texts written by non-native speakers
of English. The dataset consists of student-written sentences in
their original and revised versions with teacher feedback provided
for the errors. Feedback appears both in the form of open-ended
comments and error category tags. We focus on a specific error type,
namely linking adverbial (e.g. however, moreover) errors. The
dataset has been annotated for two aspects: (i) revision outcome
establishing whether the re-written student sentence was correct and
(ii) directness, indicating whether teachers provided explicitly the
correction in their feedback. This dataset allows for studies around
the characteristics of teacher feedback and how these influence
students' revision outcome. We describe the data preparation process
and we present initial statistical investigations regarding the
effect of different feedback characteristics on revision outcome.
These show that open-ended comments and mitigating expressions
appear in a higher proportion of successful revisions than
unsuccessful ones, while directness and metalinguistic terms have no
effect. Given that the use of this type of data is relatively
unexplored in natural language processing (NLP) applications, we
also report some observations and challenges when working with
feedback data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.41}
}

@InProceedings{evert-EtAl:2020:LREC,
  author    = {Evert, Stefan  and  Harlamov, Oleg  and  Heinrich,
Philipp  and  Banski, Piotr},
  title     = {Corpus Query Lingua Franca part II: Ontology},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3346--3352},
  abstract  = {The present paper outlines the projected second part
of the Corpus Query Lingua Franca (CQLF) family of standards: CQLF
Ontology, which is currently in the process of standardization at
the International Standards Organization (ISO), in its Technical
Committee 37, Subcommittee 4 (TC37SC4) and its national mirrors. The
first part of the family, ISO 24623-1 (henceforth CQLF Metamodel),
was successfully adopted as an international standard at the
beginning of 2018. The present paper reflects the state of the CQLF
Ontology at the moment of submission for the Committee Draft ballot.
We provide a brief overview of the CQLF Metamodel, present the
assumptions and aims of the CQLF Ontology, its basic structure, and
its potential extended applications. The full ontology is expected
to emerge from a community process, starting from an initial version
created by the authors of the present paper.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.410}
}

@InProceedings{draxler-EtAl:2020:LREC,
  author    = {Draxler, Christoph  and  van den Heuvel, Henk  and
van Hessen, Arjan  and  Calamai, Silvia  and  Corti, Louise},
  title     = {A CLARIN Transcription Portal for Interview Data},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3353--3359},
  abstract  = {In this paper we present a first version of a
transcription portal for audio files based on automatic speech
recognition (ASR) in various languages. The portal is implemented in
the CLARIN resources research network and intended for use by non-
technical scholars. We explain the background and interdisciplinary
nature of interview data, the perks and quirks of using ASR for
transcribing the audio in a research context, the dos and don’ts for
optimal use of the portal, and future developments foreseen. The
portal is promoted in a range of workshops, but there are a number
of challenges that have to be met. These challenges concern privacy
issues, ASR quality, and cost, amongst others.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.411}
}

@InProceedings{petasis-tsekouras:2020:LREC,
  author    = {Petasis, Georgios  and  Tsekouras, Leonidas},
  title     = {Ellogon Casual Annotation Infrastructure},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3360--3365},
  abstract  = {This paper presents a new annotation paradigm, casual
annotation, along with a proposed architecture and a reference
implementation, the Ellogon Casual Annotation Tool, which implements
this paradigm and architecture. The novel aspects of the proposed
paradigm originate from the vision to tightly integrate annotation
with the casual, everyday activities of users. Annotating in a less
"controlled" environment, and removing the bottleneck of selecting
content and importing it to annotation infrastructures, casual
annotation provides the ability to vastly increase the content that
can be annotated and ease the annotation process through automatic
pre-training. The proposed paradigm, architecture and reference
implementation has been evaluated for more than two years on an
annotation task related to sentiment analysis. Evaluation results
suggest that, at least for this annotation task, there is a huge
improvement in productivity after casual annotation adoption, in
comparison to the more traditional annotation paradigms followed in
```

the early stages of the annotation task.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.412}
}

@InProceedings{rehm-EtAl:2020:LREC2,
  author   = {Rehm, Georg and Berger, Maria and Elsholz, Ela and Hegele, Stefanie and Kintzel, Florian and Marheinecke, Katrin and Piperidis, Stelios and Deligiannis, Miltos and Galanis, Dimitris and Gkirtzou, Katerina and Labropoulou, Penny and Bontcheva, Kalina and Jones, David and Roberts, Ian and Hajic, Jan and Hamrlová, Jana and Kačena, Lukáš and Choukri, Khalid and Arranz, Victoria and Vasiļjevs, Andrejs and Anvari, Orians and Lagzdiņš, Andis and Meļņika, Jūlija and Backfried, Gerhard and Dikici, Erinç and Janosik, Miroslav and Prinz, Katja and Prinz, Christoph and Stampler, Severin and Thomas-Aniola, Dorothea and Gomez-Perez, Jose Manuel and Garcia Silva, Andres and Berrío, Christian and Germann, Ulrich and Renals, Steve and Klejch, Ondrej},
  title   = {European Language Grid: An Overview},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3366--3380},
  abstract = {With 24 official EU and many additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies (LTs). European LT business is dominated by hundreds of SMEs and a few large players. Many are world-class, with technologies that outperform the global players. However, European LT business is also fragmented -- by nation states, languages, verticals and sectors, significantly holding back its impact. The European Language Grid (ELG) project addresses this fragmentation by establishing the ELG as the primary platform for LT in Europe. The ELG is a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data sets and resources. Once fully operational, it will enable the commercial and non-commercial European LT community to deposit and upload their technologies and data sets into the ELG, to deploy them through the grid, and to connect with other resources. The ELG will boost the Multilingual Digital Single Market towards a thriving European LT community, creating new jobs and opportunities. Furthermore, the ELG project organises two open calls for up to 20 pilot projects. It also sets up 32 national competence centres and the European LT Council for outreach and coordination purposes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.413}
}

@InProceedings{vasijevs-EtAl:2020:LREC,
  author   = {Vasiļjevs, Andrejs and Skadina, Inguna and Samite, Indra and Kauliņš, Kaspars and Ajausks, Ēriks and

Meļņika, Jūlija and Bērziņš, Aivars},
  title   = {The Competitiveness Analysis of the European Language Technology Market},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3381--3389},
  abstract = {This paper presents the key results of a study on the global competitiveness of the European Language Technology market for three areas – Machine Translation, speech technology, and cross-lingual search. EU competitiveness is analyzed in comparison to North America and Asia. The study focuses on seven dimensions (research, innovations, investments, market dominance, industry, infrastructure, and Open Data) that have been selected to characterize the language technology market. The study concludes that while Europe still has strong positions in Research and Innovation, it lags behind North America and Asia in scaling innovations and conquering market share.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.414}
}

@InProceedings{altammami-atwell-alsalka:2020:LREC,
  author   = {Altammami, Shatha and Atwell, Eric and Alsalka, Ammar},
  title   = {Constructing a Bilingual Hadith Corpus Using a Segmentation Tool},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3390--3398},
  abstract = {This article describes the process of gathering and constructing a bilingual parallel corpus of Islamic Hadith, which is the set of narratives reporting different aspects of the prophet Muhammad's life. The corpus data is gathered from the six canonical Hadith collections using a custom segmentation tool that automatically segments and annotates the two Hadith components with 92\% accuracy. This Hadith segmenter minimises the costs of language resource creation and produces consistent results independently from previous knowledge and experiences that usually influence human annotators. The corpus includes more than 10M tokens and will be freely available via the LREC repository.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.415}
}

@InProceedings{steingrmsson-barkarson-rnlfsson:2020:LREC,
  author   = {Steingrímsson, Steinþór and Barkarson, Starkaður and Örnólfsson, Gunnar Thor},
  title   = {Facilitating Corpus Usage: Making Icelandic Corpora

More Accessible for Researchers and Language Users},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3399--3405},
  abstract = {We introduce an array of open and accessible tools to facilitate the use of the Icelandic Gigaword Corpus, in the field of Natural Language Processing as well as for students, linguists, sociologists and others benefitting from using large corpora. A KWIC engine, powered by the Swedish Korp tool is adapted to the specifics of the corpus. An n-gram viewer, highly customizable to suit different needs, allows users to study word usage throughout the period of our text collection. A frequency dictionary provides much sought after information about word frequency statistics, computed for each subcorpus as well as aggregate, disambiguating homographs based on their respective lemmas and morphosyntactic tags. Furthermore, we provide n-grams based on the corpus, and a variety of pre-trained word embeddings models, based on word2vec, GloVe, fastText and ELMo. For three of the model types, multiple word embedding models are available trained with different algorithms and using either lemmatised or unlemmatised texts.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.416}
}

@InProceedings{dejong-EtAl:2020:LREC,
  author   = {de Jong, Franciska and Maegaard, Bente and Fišer, Darja and van Uytvanck, Dieter and Witt, Andreas},
  title   = {Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3406--3413},
  abstract = {CLARIN is a European Research Infrastructure providing access to language resources and technologies for researchers in the humanities and social sciences. It supports the use and study of language data in general and aims to increase the potential for comparative research of cultural and societal phenomena across the boundaries of languages and disciplines, all in line with the European agenda for Open Science. Data infrastructures such as CLARIN have recently embarked on the emerging frameworks for the federation of infrastructural services, such as the European Open Science Cloud and the integration of services resulting from multidisciplinary collaboration in federated services for the wider SSH domain. In this paper we describe the interoperability requirements that arise through the existing ambitions and the emerging frameworks. The interoperability theme will be addressed at several levels, including organisation and ecosystem, design of

workflow services, data curation, performance measurement and collaboration.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.417}
}

@InProceedings{nikulsdttir-EtAl:2020:LREC,
  author   = {Nikulásdóttir, Anna and Guðnason, Jón and Ingason, Anton Karl and Loftsson, Hrafn and Rögnvaldsson, Eiríkur and Sigurðsson, Einar Freyr and Steingrímsson, Steinþór},
  title   = {Language Technology Programme for Icelandic 2019-2023},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3414--3422},
  abstract = {In this paper, we describe a new national language technology programme for Icelandic. The programme, which spans a period of five years, aims at making Icelandic usable in communication and interactions in the digital world, by developing accessible, open-source language resources and software. The research and development work within the programme is carried out by a consortium of universities, institutions, and private companies, with a strong emphasis on cooperation between academia and industries. Five core projects will be the main content of the programme: language resources, speech recognition, speech synthesis, machine translation, and spell and grammar checking. We also describe other national language technology programmes and give an overview over the history of language technology in Iceland.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.418}
}

@InProceedings{kamocki-witt:2020:LREC,
  author   = {Kamocki, Pawel and Witt, Andreas},
  title   = {Privacy by Design and Language Resources},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3423--3427},
  abstract = {Privacy by Design (also referred to as Data Protection by Design) is an approach in which solutions and mechanisms addressing privacy and data protection are embedded through the entire project lifecycle, from the early design stage, rather than just added as an additional lawyer to the final product. Formulated in the 1990 by the Privacy Commissionner of Ontario, the principle of Privacy by Design has been discussed by institutions and policymakers on both sides of the Atlantic, and mentioned already in the 1995 EU Data Protection Directive (95/46/EC). More

recently, Privacy by Design was introduced as one of the requirements of the General Data Protection Regulation (GDPR), obliging data controllers to define and adopt, already at the conception phase, appropriate measures and safeguards to implement data protection principles and protect the rights of the data subject. Failing to meet this obligation may result in a hefty fine, as it was the case in the Uniontrad decision by the French Data Protection Authority (CNIL). The ambition of the proposed paper is to analyse the practical meaning of Privacy by Design in the context of Language Resources, and propose measures and safeguards that can be implemented by the community to ensure respect of this principle.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.419}
}

@InProceedings{nagata-inui-ishikawa:2020:LREC,
  author   = {Nagata, Ryo and Inui, Kentaro and Ishikawa, Shin'ichiro},
  title   = {Creating Corpora for Research in Feedback Comment Generation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {340--345},
  abstract = {In this paper, we report on datasets that we created for research in feedback comment generation – a task of automatically generating feedback comments such as a hint or an explanatory note for writing learning. There has been almost no such corpus open to the public and accordingly there has been a very limited amount of work on this task. In this paper, we first discuss the principle and guidelines for feedback comment annotation. Then, we describe two corpora that we have manually annotated with feedback comments (approximately 50,000 general comments and 6,700 on preposition use). A part of the annotation results is now available on the web, which will facilitate research in feedback comment generation},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.42}
}

@InProceedings{labropoulou-EtAl:2020:LREC,
  author   = {Labropoulou, Penny and Gkirtzou, Katerina and Gavriilidou, Maria and Deligiannis, Miltos and Galanis, Dimitris and Piperidis, Stelios and Rehm, Georg and Berger, Maria and Mapelli, Valérie and Rigault, Michael and Arranz, Victoria and Choukri, Khalid and Backfried, Gerhard and Gomez-Perez, Jose Manuel and Garcia-Silva, Andres},
  title   = {Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3428--3437},
  abstract  = {The current scientific and technological landscape is
characterised by the increasing availability of data resources and
processing tools and services. In this setting, metadata have
emerged as a key factor facilitating management, sharing and usage
of such digital assets. In this paper we present ELG-SHARE, a rich
metadata schema catering for the description of Language Resources
and Technologies (processing and generation services and tools,
models, corpora, term lists, etc.), as well as related entities
(e.g., organizations, projects, supporting documents, etc.). The
schema powers the European Language Grid platform that aims to be
the primary hub and marketplace for industry-relevant Language
Technology in Europe. ELG-SHARE has been based on various metadata
schemas, vocabularies, and ontologies, as well as related
recommendations and guidelines.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.420}
}

@InProceedings{jaquette-cieri-dipersio:2020:LREC,
  author    = {Jaquette, Daniel  and  Cieri, Christopher  and
DiPersio, Denise},
  title     = {Related Works in the Linguistic Data Consortium
Catalog},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3438--3442},
  abstract  = {Defining relations between language resources
provides an archive with the ability to better serve its users. This
paper covers the development and implementation of a Related Works
addition to the Linguistic Data Consortium's (LDC) catalog. The
authors go step-by-step through the development of the Related Works
schema, implementation of the software and database changes, and
data entry of the relations. The Related Work schema involved
developing of a set of controlled terms for relations based on
previous work and other schema. Software and database changes
consisted of both front and back end interface additions, along with
modification and additions to the LDC Catalog database tables. Data
entry consisted of two parts: seed data from previous work and 2019
language resources, and ongoing legacy population. Previous work in
this area is discussed as well as overview information about the LDC
Catalog. A list of the full LDC Related Works terms is included with
brief explanations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.421}
}

@InProceedings{smal-EtAl:2020:LREC,
```

```
  author    = {Smal, Lilli  and  Lösch, Andrea  and  van Genabith,
Josef  and  Giagkou, Maria  and  Declerck, Thierry  and  Busemann,
Stephan},
  title     = {Language Data Sharing in European Public Services –
Overcoming Obstacles and Creating Sustainable Data Sharing
Infrastructures},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3443--3448},
  abstract  = {Data is key in training modern language technologies.
In this paper, we summarise the findings of the first pan-European
study on obstacles to sharing language data across 29 EU Member
States and CEF-affiliated countries carried out under the ELRC White
Paper action on Sustainable Language Data Sharing to Support
Language Equality in Multilingual Europe. Why Language Data Matters.
We present the methodology of the study, the obstacles identified
and report on recommendations on how to overcome those. The
obstacles are classified into (1) lack of appreciation of the value
of language data, (2) structural challenges, (3) disposition towards
CAT tools and lack of digital skills, (4) inadequate language data
management practices, (5) limited access to outsourced translations,
and (6) legal concerns. Recommendations are grouped into addressing
the European/national policy level, and the organisational/
institutional level.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.422}
}

@InProceedings{cieri-EtAl:2020:LREC,
  author    = {Cieri, Christopher  and  Fiumara, James  and
Strassel, Stephanie  and  Wright, Jonathan  and  DiPersio, Denise
and  Liberman, Mark},
  title     = {A Progress Report on Activities at the Linguistic
Data Consortium Benefitting the LREC Community},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3449--3456},
  abstract  = {This latest in a series of Linguistic Data Consortium
(LDC) progress reports to the LREC community does not describe any
single language resource, evaluation campaign or technology but
sketches the activities, since the last report, of a data center
devoted to supporting the work of LREC attendees among other
research communities. Specifically, we describe 96 new corpora
released in 2018-2020 to date, a new technology evaluation campaign,
ongoing activities to support multiple common task human language
technology programs, and innovations to advance the methodology of
language data collection and annotation.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.423}
}

@InProceedings{lyding-knig-pretti:2020:LREC,
  author    = {Lyding, Verena  and  König, Alexander  and  Pretti,
Monica},
  title     = {Digital Language Infrastructures – Documenting
Language Actors},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3457--3462},
  abstract  = {The major European language infrastructure
initiatives like CLARIN (Hinrichs and Krauwer, 2014), DARIAH (Edmond
et al., 2017) or Europeana (Europeana Foundation, 2015) have been
built by focusing in the first place on institutions of larger
scale, like specialized research departments and larger official
units like national libraries, etc. However, besides these principal
players also a large number of smaller language actors could
contribute to and benefit from language infrastructures. Especially
since these smaller institutions, like local libraries, archives and
publishers, often collect, manage and host language resources of
particular value for their geographical and cultural region, it
seems highly relevant to find ways of engaging and connecting them
to existing European infrastructure initiatives. In this article, we
first highlight the need for reaching out to smaller local language
actors and discuss challenges related to this ambition. Then we
present the first step in how this objective was approached within a
local language infrastructure project, namely by means of a
structured documentation of the local language actors landscape in
South Tyrol. We describe how the documentation efforts were
structured and organized, and what tool we have set up to distribute
the collected data online, by adapting existing CLARIN solutions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.424}
}

@InProceedings{mollberg-EtAl:2020:LREC,
  author    = {Mollberg, David Erik  and  Jónsson, Ólafur Helgi  and
Þorsteinsdóttir, Sunneva  and  Steingrímsson, Steinþór  and
Magnúsdóttir, Eydís Huld  and  Gudnason, Jon},
  title     = {Samrómur: Crowd-sourcing Data Collection for
Icelandic Speech Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3463--3467},
  abstract  = {This contribution describes an ongoing project of
speech data collection, using the web application Samrómur which is
```

```
built upon Common Voice, Mozilla Foundation's web platform for open-source voice collection. The goal of the project is to build a large-scale speech corpus for Automatic Speech Recognition (ASR) for Icelandic. Upon completion, Samrómur will be the largest open speech corpus for Icelandic collected from the public domain. We discuss the methods used for the crowd-sourcing effort and show the importance of marketing and good media coverage when launching a crowd-sourcing campaign. Preliminary results exceed our expectations, and in one month we collected data that we had estimated would take three months to obtain. Furthermore, our initial dataset of around 45 thousand utterances has good demographic coverage, is gender-balanced and with proper age distribution. We also report on the task of validating the recordings, which we have not promoted, but have had numerous hours invested by volunteers.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.425}
}

@InProceedings{biswas-EtAl:2020:LREC,
  author    = {Biswas, Astik  and  Yilmaz, Emre  and  De Wet, Febe and  Van der westhuizen, Ewald  and  Niesler, Thomas},
  title     = {Semi-supervised Development of ASR Systems for Multilingual Code-switched Speech in Under-resourced Languages},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3468--3474},
  abstract  = {This paper reports on the semi-supervised development of acoustic and language models for under-resourced, code-switched speech in five South African languages. Two approaches are considered. The first constructs four separate bilingual automatic speech recognisers (ASRs) corresponding to four different language pairs between which speakers switch frequently. The second uses a single, unified, five-lingual ASR system that represents all the languages (English, isiZulu, isiXhosa, Setswana and Sesotho). We evaluate the effectiveness of these two approaches when used to add additional data to our extremely sparse training sets. Results indicate that batch-wise semi-supervised training yields better results than a non-batch-wise approach. Furthermore, while the separate bilingual systems achieved better recognition performance than the unified system, they benefited more from pseudolabels generated by the five-lingual system than from those generated by the bilingual systems.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.426}
}

@InProceedings{mersinias-afantenos-chalkiadakis:2020:LREC,
  author    = {Mersinias, Michail  and  Afantenos, Stergos  and  Chalkiadakis, Georgios},
  title     = {CLFD: A Novel Vectorization Technique and Its Application in Fake News Detection},
```

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3475--3483},
  abstract  = {In recent years, fake news detection has been an
emerging research area. In this paper, we put forward a novel
statistical approach for the generation of feature vectors to
describe a document. Our so-called class label frequency distance
(clfd), is shown experimentally to provide an effective way for
boosting the performance of machine learning methods. Specifically,
our experiments, carried out in the fake news detection domain,
verify that efficient traditional machine learning methods that use
our vectorization approach, consistently outperform deep learning
methods that use word embeddings for small and medium sized
datasets, while the results are comparable for large datasets. In
addition, we demonstrate that a novel hybrid method that utilizes
both a clfd-boosted logistic regression classifier and a deep
learning one, clearly outperforms deep learning methods even in
large datasets.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.427}
}

@InProceedings{qasmi-EtAl:2020:LREC,
  author    = {Qasmi, Namoos Hayat  and  Zia, Haris Bin  and  Athar,
Awais  and  Raza, Agha Ali},
  title     = {SimplifyUR: Unsupervised Lexical Text Simplification
for Urdu},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3484--3489},
  abstract  = {This paper presents the first attempt at Automatic
Text Simplification (ATS) for Urdu, the language of 170 million
people worldwide. Being a low-resource language in terms of standard
linguistic resources, recent text simplification approaches that
rely on manually crafted simplified corpora or lexicons such as
WordNet are not applicable to Urdu. Urdu is a morphologically rich
language that requires unique considerations such as proper handling
of inflectional case and honorifics. We present an unsupervised
method for lexical simplification of complex Urdu text. Our method
only requires plain Urdu text and makes use of word embeddings
together with a set of morphological features to generate
simplifications. Our system achieves a BLEU score of 80.15 and SARI
score of 42.02 upon automatic evaluation on manually crafted
simplified corpora. We also report results for human evaluations for
correctness, grammaticality, meaning-preservation and simplicity of
the output. Our code and corpus are publicly available to make our
results reproducible.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.428}
}

@InProceedings{moon-okazaki:2020:LREC,
  author    = {Moon, Sangwhan  and  Okazaki, Naoaki},
  title     = {Jamo Pair Encoding: Subcharacter Representation-based
Extreme Korean Vocabulary Compression for Efficient Subword
Tokenization},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3490--3497},
  abstract  = {In the context of multilingual language model pre-
training, vocabulary size for languages with a broad set of
potential characters is an unsolved problem. We propose two
algorithms applicable in any unsupervised multilingual pre-training
task, increasing the elasticity of budget required for building the
vocabulary in Byte-Pair Encoding inspired tokenizers, significantly
reducing the cost of supporting Korean in a multilingual model.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.429}
}

@InProceedings{gran-alfter-schneider:2020:LREC,
  author    = {Graën, Johannes  and  Alfter, David  and  Schneider,
Gerold},
  title     = {Using Multilingual Resources to Evaluate CEFRLex for
Learner Applications},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {346--355},
  abstract  = {The Common European Framework of Reference for
Languages (CEFR) defines six levels of learner proficiency, and
links them to particular communicative abilities. The CEFRLex
project aims at compiling lexical resources that link single words
and multi-word expressions to particular CEFR levels. The resources
are thought to reflect second language learner needs as they are
compiled from CEFR-graded textbooks and other learner-directed
texts. In this work, we investigate the applicability of CEFRLex
resources for building language learning applications. Our main
concerns were that vocabulary in language learning materials might
be sparse, i.e. that not all vocabulary items that belong to a
particular level would also occur in materials for that level, and,
on the other hand, that vocabulary items might be used on lower-
level materials if required by the topic (e.g. with a simpler
paraphrasing or translation). Our results indicate that the English
CEFRLex resource is in accordance with external resources that we
jointly employ as gold standard. Together with other values obtained
```

from monolingual and parallel corpora, we can indicate which entries need to be adjusted to obtain values that are even more in line with this gold standard. We expect that this finding also holds for the other languages},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.43}
}

@InProceedings{sigurbergsson-derczynski:2020:LREC,
  author   = {Sigurbergsson, Gudbjartur Ingi and Derczynski, Leon},
  title   = {Offensive Language and Hate Speech Detection for Danish},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3498--3508},
  abstract = {The presence of offensive language on social media platforms and the implications this poses is becoming a major concern in modern society. Given the enormous amount of content created every day, automatic methods are required to detect and deal with this type of content. Until now, most of the research has focused on solving the problem for the English language, while the problem is multilingual. We construct a Danish dataset DKhate containing user-generated comments from various social media platforms, and to our knowledge, the first of its kind, annotated for various types and target of offensive language. We develop four automatic classification systems, each designed to work for both the English and the Danish language. In the detection of offensive language in English, the best performing system achieves a macro averaged F1-score of 0.74, and the best performing system for Danish achieves a macro averaged F1-score of 0.70. In the detection of whether or not an offensive post is targeted, the best performing system for English achieves a macro averaged F1-score of 0.62, while the best performing system for Danish achieves a macro averaged F1-score of 0.73. Finally, in the detection of the target type in a targeted offensive post, the best performing system for English achieves a macro averaged F1-score of 0.56, and the best performing system for Danish achieves a macro averaged F1-score of 0.63. Our work for both the English and the Danish language captures the type and targets of offensive language, and present automatic methods for detecting different kinds of offensive language such as hate speech and cyberbullying.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.430}
}

@InProceedings{yong-timponitorrent:2020:LREC,
  author   = {Yong, Zheng Xin and Timponi Torrent, Tiago},
  title   = {Semi-supervised Deep Embedded Clustering with Anomaly Detection for Semantic Frame Induction},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3509--3519},
  abstract  = {Although FrameNet is recognized as one of the most
fine-grained lexical databases, its coverage of lexical units is
still limited. To tackle this issue, we propose a two-step frame
induction process: for a set of lexical units not yet present in
Berkeley FrameNet data release 1.7, first remove those that cannot
fit into any existing semantic frame in FrameNet; then, assign the
remaining lexical units to their correct frames. We also present the
Semi-supervised Deep Embedded Clustering with Anomaly Detection
(SDEC-AD) model—an algorithm that maps high-dimensional
contextualized vector representations of lexical units to a low-
dimensional latent space for better frame prediction and uses
reconstruction error to identify lexical units that cannot evoke
frames in FrameNet. SDEC-AD outperforms the state-of-the-art methods
in both steps of the frame induction process. Empirical results also
show that definitions provide contextual information for
representing and characterizing the frame membership of lexical
units.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.431}
}

@InProceedings{tambi-kale-king:2020:LREC,
  author    = {Tambi, Ritiz  and  Kale, Ajinkya  and  King, Tracy
Holloway},
  title     = {Search Query Language Identification Using Weak
Labeling},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3520--3527},
  abstract  = {Language identification is a well-known task for
natural language documents. In this paper we explore search query
language identification which is usually the first task before any
other query understanding. Without loss of generalization, we run
our experiments on the Adobe Stock search engine. Even though the
domain is relatively generic because Adobe Stock queries cover a
broad range of objects and concepts, out-of-the-box language
identifiers do not perform well due to the extremely short text
found in queries. Unlike other well-studied supervised approaches
for this task, we examine a practical approach for the cold start
problem for automatically getting large-scale query-language pairs
for training. We describe the process of creating weak-labeled
training data and then human-annotated evaluation data for the
search query language identification task. The effectiveness of this
technique is demonstrated by training a gradient boosting model for
language classification given a query. We out-perform the open
domain text model baselines by a large margin.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.432}
}

@InProceedings{sahala-EtAl:2020:LREC1,
  author    = {Sahala, Aleksi  and  Silfverberg, Miikka  and  Arppe,
Antti  and  Lindén, Krister},
  title     = {Automated Phonological Transcription of Akkadian
Cuneiform Text},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3528--3534},
  abstract  = {Akkadian was an East-Semitic language spoken in
ancient Mesopotamia. The language is attested on hundreds of
thousands of cuneiform clay tablets. Several Akkadian text corpora
contain only the transliterated text. In this paper, we investigate
automated phonological transcription of the transliterated corpora.
The phonological transcription provides a linguistically appealing
form to represent Akkadian, because the transcription is normalized
according to the grammatical description of a given dialect and
explicitly shows the Akkadian renderings for Sumerian logograms.
Because cuneiform text does not mark the inflection for logograms,
the inflected form needs to be inferred from the sentence context.
To the best of our knowledge, this is the first documented attempt
to automatically transcribe Akkadian. Using a context-aware neural
network model, we are able to automatically transcribe syllabic
tokens at near human performance with 96\% recall @ 3, while the
logogram transcription remains more challenging at 82\% recall @
3.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.433}
}

@InProceedings{barancikova-bojar:2020:LREC,
  author    = {Barancikova, Petra  and  Bojar, Ondřej},
  title     = {COSTRA 1.0: A Dataset of Complex Sentence
Transformations},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3535--3541},
  abstract  = {We present COSTRA 1.0, a dataset of complex sentence
transformations. The dataset is intended for the study of sentence-
level embeddings beyond simple word alternations or standard
paraphrasing. This first version of the dataset is limited to
sentences in Czech but the construction method is universal and we
plan to use it also for other languages. The dataset consist of
4,262 unique sentences with average length of 10 words, illustrating
15 types of modifications such as simplification, generalization, or
```

formal and informal language variation. The hope is that with this dataset, we should be able to test semantic properties of sentence embeddings and perhaps even to find some topologically interesting ``skeleton'' in the sentence embedding space. A preliminary analysis using LASER, multi-purpose multi-lingual sentence embeddings suggests that the LASER space does not exhibit the desired properties.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.434}
}

@InProceedings{correia-trancoso-raj:2020:LREC,
  author   = {Correia, Joana and Trancoso, Isabel and Raj, Bhiksha},
  title   = {Automatic In-the-wild Dataset Annotation with Deep Generalized Multiple Instance Learning},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3542--3550},
  abstract = {The automation of the diagnosis and monitoring of speech affecting diseases in real life situations, such as Depression or Parkinson's disease, depends on the existence of rich and large datasets that resemble real life conditions, such as those collected from in-the-wild multimedia repositories like YouTube. However, the cost of manually labeling these large datasets can be prohibitive. In this work, we propose to overcome this problem by automating the annotation process, without any requirements for human intervention. We formulate the annotation problem as a Multiple Instance Learning (MIL) problem, and propose a novel solution that is based on end-to-end differentiable neural networks. Our solution has the additional advantage of generalizing the MIL framework to more scenarios where the data is stil organized in bags but does not meet the MIL bag label conditions. We demonstrate the performance of the proposed method in labeling the in-the-Wild Speech Medical (WSM) Corpus, using simple textual cues extracted from videos and their metadata. Furthermore we show what is the contribution of each type of textual cues for the final model performance, as well as study the influence of the size of the bags of instances in determining the difficulty of the learning problem},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.435}
}

@InProceedings{strbel-clematide-volk:2020:LREC,
  author   = {Ströbel, Phillip Benjamin and Clematide, Simon and Volk, Martin},
  title   = {How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3551--3559},
  abstract  = {Recent advances in Optical Character Recognition
(OCR) and Handwritten Text Recognition (HTR) have led to more
accurate textrecognition of historical documents. The Digital
Humanities heavily profit from these developments, but they still
struggle whenchoosing from the plethora of OCR systems available on
the one hand and when defining workflows for their projects on the
other hand.In this work, we present our approach to build a ground
truth for a historical German-language newspaper published in black
letter. Wealso report how we used it to systematically evaluate the
performance of different OCR engines. Additionally, we used this
ground truthto make an informed estimate as to how much data is
necessary to achieve high-quality OCR results. The outcomes of our
experimentsshow that HTR architectures can successfully recognise
black letter text and that a ground truth size of 50 newspaper pages
suffices toachieve good OCR accuracy. Moreover, our models perform
equally well on data they have not seen during training, which means
thatadditional manual correction for diverging data is
superfluous.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.436}
}

@InProceedings{jungmaier-kassner-roth:2020:LREC,
  author    = {Jungmaier, Jakob  and  Kassner, Nora  and  Roth,
Benjamin},
  title     = {Dirichlet-Smoothed Word Embeddings for Low-Resource
Settings},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3560--3565},
  abstract  = {Nowadays, classical count-based word embeddings using
positive pointwise mutual information (PPMI) weighted co-occurrence
matrices have been widely superseded by machine-learning-based
methods like word2vec and GloVe. But these methods are usually
applied using very large amounts of text data. In many cases,
however, there is not much text data available, for example for
specific domains or low-resource languages. This paper revisits PPMI
by adding Dirichlet smoothing to correct its bias towards rare
words. We evaluate on standard word similarity data sets and compare
to word2vec and the recent state of the art for low-resource
settings: Positive and Unlabeled (PU) Learning for word embeddings.
The proposed method outperforms PU-Learning for low-resource
settings and obtains competitive results for Maltese and
Luxembourgish.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.437}
}

@InProceedings{monteiro-alam-falk:2020:LREC,
```

```
  author    = {Monteiro, Joao  and  Alam, Md Jahangir  and  Falk,
Tiago},
  title     = {On The Performance of Time-Pooling Strategies for
End-to-End Spoken Language Identification},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3566--3572},
  abstract  = {Automatic speech processing applications often have
to deal with the problem of aggregating local descriptors (i.e.,
representations of input speech data corresponding to specific
portions across the time dimension) and turning them into a single
fixed-dimension representation, known as global descriptor, on top
of which downstream classification tasks can be performed. In this
paper, we provide an empirical assessment of different time pooling
strategies when used with state-of-the-art representation learning
models. In particular, insights are provided as to when it is
suitable to use simple statistics of local descriptors or when more
sophisticated approaches are needed. Here, language identification
is used as a case study and a database containing ten oriental
languages under varying test conditions (short-duration test
recordings, confusing languages, unseen languages) is used.
Experiments are performed with classifiers trained on top of global
descriptors to provide insights on open-set evaluation performance
and show that appropriate selection of such pooling strategies yield
embeddings able to outperform well-known benchmark systems as well
as previously results based on attention only.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.438}
}

@InProceedings{hoyaquecedo-maximilian-yangarber:2020:LREC,
  author    = {Hoya Quecedo, José María  and  Maximilian, Koppatz
and  Yangarber, Roman},
  title     = {Neural Disambiguation of Lemma and Part of Speech in
Morphologically Rich Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3573--3582},
  abstract  = {We consider the problem of disambiguating the lemma
and part of speech of ambiguous words in morphologically rich
languages. We propose a method for disambiguating ambiguous words in
context, using a large un-annotated corpus of text, and a
morphological analyser—with no manual disambiguation or data
annotation. We assume that the morphological analyser produces
multiple analyses for ambiguous words. The idea is to train
recurrent neural networks on the output that the morphological
analyser produces for unambiguous words. We present performance on
```

```
POS and lemma disambiguation that reaches or surpasses the state of
the art—including supervised models—using no manually annotated
data. We evaluate the method on several morphologically rich
languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.439}
}

@InProceedings{platte-EtAl:2020:LREC,
  author    = {Platte, Benny  and  Platte, Anett  and  Roschke,
Christian  and  Thomanek, Rico  and  Rolletschke, Thony  and
Zimmer, Frank  and  Ritter, Marc},
  title     = {Immersive Language Exploration with Object
Recognition and Augmented Reality},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {356--362},
  abstract  = {The use of Augmented Reality (AR) in teaching and
learning contexts for language is still young. The ideas are
endless, the concrete educational offers available emerge only
gradually. Educational opportunities that were unthinkable a few
years ago are now feasible. We present a concrete realization: an
executable application for mobile devices with which users can
explore their environment interactively in different languages. The
software recognizes up to 1000 objects in the user's environment
using a deep learning method based on Convolutional Neural Networks
and names this objects accordingly. Using Augmented Reality the
objects are superimposed with 3D information in different languages.
By switching the languages, the user is able to interactively
discover his surrounding everyday items in all languages. The
application is available as Open Source.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.44}
}

@InProceedings{zhao-gilman:2020:LREC,
  author    = {Zhao, Jiawei  and  Gilman, Andrew},
  title     = {Non-Linearity in Mapping Based Cross-Lingual Word
Embeddings},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3583--3589},
  abstract  = {Recent works on cross-lingual word embeddings have
been mainly focused on linear-mapping-based approaches, where pre-
trained word embeddings are mapped into a shared vector space using
a linear transformation. However, there is a limitation in such
approaches--they follow a key assumption: words with similar
meanings share similar geometric arrangements between their
```

```
monolingual word embeddings, which suggest that there is a linear
relationship between languages. However, such assumption may not
hold for all language pairs across all semantic concepts. We
investigate whether non-linear mappings can better describe the
relationship between different languages by utilising kernel
Canonical Correlation Analysis (KCCA). Experimental results on five
language pairs show an improvement over current state-of-art results
in both supervised and self-learning scenarios, confirming that non-
linear mapping is a better way to describe the relationship between
languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.440}
}

@InProceedings{beilharz-EtAl:2020:LREC,
  author    = {Beilharz, Benjamin  and  Sun, Xin  and  Karimova,
Sariya  and  Riezler, Stefan},
  title     = {LibriVoxDeEn: A Corpus for German-to-English Speech
Translation and German Speech Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3590--3594},
  abstract  = {We present a corpus of sentence-aligned triples of
German audio, German text, and English translation, based on German
audio books. The speech translation data consist of 110 hours of
audio material aligned to over 50k parallel sentences. An even
larger dataset comprising 547 hours of German speech aligned to
German text is available for speech recognition. The audio data is
read speech and thus low in disfluencies. The quality of audio and
sentence alignments has been checked by a manual evaluation, showing
that speech alignment quality is in general very high. The sentence
alignment quality is comparable to well-used parallel translation
data and can be adjusted by cutoffs on the automatic alignment
score. To our knowledge, this corpus is to date the largest resource
for German speech recognition and for end-to-end German-to-English
speech translation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.441}
}

@InProceedings{ghaddar-langlais:2020:LREC,
  author    = {Ghaddar, Abbas  and  Langlais, Phillippe},
  title     = {SEDAR: a Large Scale French-English Financial Domain
Parallel Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3595--3602},
  abstract  = {This paper describes the acquisition, preprocessing
```

and characteristics of SEDAR, a large scale English-French parallel corpus for the financial domain. Our extensive experiments on machine translation show that SEDAR is essential to obtain good performance on finance. We observe a large gain in the performance of machine translation systems trained on SEDAR when tested on finance, which makes SEDAR suitable to study domain adaptation for neural machine translation. The first release of the corpus comprises 8.6 million high quality sentence pairs that are publicly available for research at https://github.com/autorite/sedar-bitext.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.442}
}

@InProceedings{morishita-suzuki-nagata:2020:LREC,
  author   = {Morishita, Makoto and Suzuki, Jun and Nagata, Masaaki},
  title   = {JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3603--3609},
  abstract = {Recent machine translation algorithms mainly rely on parallel corpora. However, since the availability of parallel corpora remains limited, only some resource-rich language pairs can benefit from them. We constructed a parallel corpus for English-Japanese, for which the amount of publicly available parallel corpora is still limited. We constructed the parallel corpus by broadly crawling the web and automatically aligning parallel sentences. Our collected corpus, called JParaCrawl, amassed over 8.7 million sentence pairs. We show how it includes a broader range of domains and how a neural machine translation model trained with it works as a good pre-trained model for fine-tuning specific domains. The pre-training and fine-tuning approaches achieved or surpassed performance comparable to model training from the initial state and reduced the training time. Additionally, we trained the model with an in-domain dataset and JParaCrawl to show how we achieved the best performance with them. JParaCrawl and the pre-trained models are freely available online for research purposes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.443}
}

@InProceedings{choudhary-rao-rohilla:2020:LREC,
  author   = {Choudhary, Himanshu and Rao, Shivansh and Rohilla, Rajesh},
  title   = {Neural Machine Translation for Low-Resourced Indian Languages},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3610--3615},
  abstract  = {A large number of significant assets are available
online in English, which is frequently translated into native
languages to ease the information sharing among local people who are
not much familiar with English. However, manual translation is a
very tedious, costly, and time-taking process. To this end, machine
translation is an effective approach to convert text to a different
language without any human involvement. Neural machine translation
(NMT) is one of the most proficient translation techniques amongst
all existing machine translation systems. In this paper, we have
applied NMT on two of the most morphological rich Indian languages,
i.e. English-Tamil and English-Malayalam. We proposed a novel NMT
model using Multihead self-attention along with pre-trained Byte-
Pair-Encoded (BPE) and MultiBPE embeddings to develop an efficient
translation system that overcomes the OOV (Out Of Vocabulary)
problem for low resourced morphological rich Indian languages which
do not have much translation available online. We also collected
corpus from different sources, addressed the issues with these
publicly available data and refined them for further uses. We used
the BLEU score for evaluating our system performance. Experimental
results and survey confirmed that our proposed translator (24.34 and
9.78 BLEU score) outperforms Google translator (9.40 and 5.94 BLEU
score) respectively.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.444}
}

@InProceedings{mino-EtAl:2020:LREC,
  author    = {Mino, Hideya  and  Tanaka, Hideki  and  Ito, Hitoshi
and  Goto, Isao  and  Yamada, Ichiro  and  Tokunaga, Takenobu},
  title     = {Content-Equivalent Translated Parallel News Corpus
and Extension of Domain Adaptation for NMT},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3616--3622},
  abstract  = {In this paper, we deal with two problems in Japanese-
English machine translation of news articles. The first problem is
the quality of parallel corpora. Neural machine translation (NMT)
systems suffer degraded performance when trained with noisy data.
Because there is no clean Japanese-English parallel data for news
articles, we build a novel parallel news corpus consisting of
Japanese news articles translated into English in a content-
equivalent manner. This is the first content-equivalent Japanese-
English news corpus translated specifically for training NMT
systems. The second problem involves the domain-adaptation
technique. NMT systems suffer degraded performance when trained with
mixed data having different features, such as noisy data and clean
data. Though the existing methods try to overcome this problem by
using tags for distinguishing the differences between corpora, it is
```

not sufficient. We thus extend a domain-adaptation method using multi-tags to train an NMT model effectively with the clean corpus and existing parallel news corpora with some types of noise. Experimental results show that our corpus increases the translation quality, and that our domain-adaptation method is more effective for learning with the multiple types of corpora than existing domain-adaptation methods are.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.445}
}

@InProceedings{caseli-incio:2020:LREC,
  author   = {Caseli, Helena and Inácio, Marcio},
  title   = {NMT and PBSMT Error Analyses in English to Brazilian Portuguese Automatic Translations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3623--3629},
  abstract = {Machine Translation (MT) is one of the most important natural language processing applications. Independently of the applied MT approach, a MT system automatically generates an equivalent version (in some target language) of an input sentence (in some source language). Recently, a new MT approach has been proposed: neural machine translation (NMT). NMT systems have already outperformed traditional phrase-based statistical machine translation (PBSMT) systems for some pairs of languages. However, any MT approach outputs errors. In this work we present a comparative study of MT errors generated by a NMT system and a PBSMT system trained on the same English -- Brazilian Portuguese parallel corpus. This is the first study of this kind involving NMT for Brazilian Portuguese. Furthermore, the analyses and conclusions presented here point out the specific problems of NMT outputs in relation to PBSMT ones and also give lots of insights into how to implement automatic post-editing for a NMT system. Finally, the corpora annotated with MT errors generated by both PBSMT and NMT systems are also available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.446}
}

@InProceedings{shimazu-EtAl:2020:LREC,
  author   = {Shimazu, Sho and Takase, Sho and Nakazawa, Toshiaki and Okazaki, Naoaki},
  title   = {Evaluation Dataset for Zero Pronoun in Japanese to English Translation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3630--3634},

```
  abstract  = {In natural language, we often omit some words that
are easily understandable from the context. In particular, pronouns
of subject, object, and possessive cases are often omitted in
Japanese; these are known as zero pronouns. In translation from
Japanese to other languages, we need to find a correct antecedent
for each zero pronoun to generate a correct and coherent
translation. However, it is difficult for conventional automatic
evaluation metrics (e.g., BLEU) to focus on the success of zero
pronoun resolution. Therefore, we present a hand-crafted dataset to
evaluate whether translation models can resolve the zero pronoun
problems in Japanese to English translations. We manually and
statistically validate that our dataset can effectively evaluate the
correctness of the antecedents selected in translations. Through the
translation experiments using our dataset, we reveal shortcomings of
an existing context-aware neural machine translation model.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.447}
}

@InProceedings{kovcs-EtAl:2020:LREC,
  author    = {Kovács, Ádám  and  Ács, Judit  and  Kornai, Andras
and  Recski, Gábor},
  title     = {Better Together: Modern Methods Plus Traditional
Thinking in NP Alignment},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3635--3639},
  abstract  = {We study a typical intermediary task to Machine
Translation, the alignment of NPs in the bitext. After arguing that
the task remains relevant even in an end-to-end paradigm, we present
simple, dictionary- and word vector-based baselines and a BERT-based
system. Our results make clear that even state of the art systems
relying on the best end-to-end methods can be improved by bringing
in old-fashioned methods such as stopword removal, lemmatization,
and dictionaries},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.448}
}

@InProceedings{song-EtAl:2020:LREC3,
  author    = {Song, Haiyue  and  Dabre, Raj  and  Fujita, Atsushi
and  Kurohashi, Sadao},
  title     = {Coursera Corpus Mining and Multistage Fine-Tuning for
Improving Lectures Translation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3640--3649},
  abstract  = {Lectures translation is a case of spoken language
```

translation and there is a lack of publicly available parallel corpora for this purpose. To address this, we examine a framework for parallel corpus mining which is a quick and effective way to mine a parallel corpus from publicly available lectures at Coursera. Our approach determines sentence alignments, relying on machine translation and cosine similarity over continuous-space sentence representations. We also show how to use the resulting corpora in a multistage fine-tuning based domain adaptation for high-quality lectures translation. For Japanese--English lectures translation, we extracted parallel data of approximately 40,000 lines and created development and test sets through manual filtering for benchmarking translation performance. We demonstrate that the mined corpus greatly enhances the quality of translation when used in conjunction with out-of-domain parallel corpora via multistage training. This paper also suggests some guidelines to gather and clean corpora, mine parallel sentences, address noise in the mined data, and create high-quality evaluation splits. For the sake of reproducibility, we have released our code for parallel data creation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.449}
}

@InProceedings{conijn-EtAl:2020:LREC,
  author   = {Conijn, Rianne and Dux Speltz, Emily and van Zaanen, Menno and Van Waes, Luuk and Chukharev-Hudilainen, Evgeny},
  title   = {A Process-oriented Dataset of Revisions during Writing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {363--368},
  abstract = {Revision plays a major role in writing and the analysis of writing processes. Revisions can be analyzed using a product-oriented approach (focusing on a finished product, the text that has been produced) or a process-oriented approach (focusing on the process that the writer followed to generate this product). Although several language resources exist for the product-oriented approach to revisions, there are hardly any resources available yet for an in-depth analysis of the process of revisions. Therefore, we provide an extensive dataset on revisions made during writing (accessible via https://hdl.handle.net/10411/VBDYGX). This dataset is based on keystroke data and eye tracking data of 65 students from a variety of backgrounds (undergraduate and graduate English as a first language and English as a second language students) and a variety of tasks (argumentative text and academic abstract). In total, 7,120 revisions were identified in the dataset. For each revision, 18 features have been manually annotated and 31 features have been automatically extracted. As a case study, we show two potential use cases of the dataset. In addition, future uses of the dataset are described.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.45}

}

@InProceedings{defauw-EtAl:2020:LREC,
  author   = {Defauw, Arne and Vanallemeersch, Tom and Van Winckel, Koen and Szoc, Sara and Van den Bogaert, Joachim},
  title   = {Being Generous with Sub-Words towards Small NMT Children},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3650--3656},
  abstract = {In the context of under-resourced neural machine translation (NMT), transfer learning from an NMT model trained on a high resource language pair, or from a multilingual NMT (M-NMT) model, has been shown to boost performance to a large extent. In this paper, we focus on so-called cold start transfer learning from an M-NMT model, which means that the parent model is not trained on any of the child data. Such a set-up enables quick adaptation of M-NMT models to new languages. We investigate the effectiveness of cold start transfer learning from a many-to-many M-NMT model to an under-resourced child. We show that sufficiently large sub-word vocabularies should be used for transfer learning to be effective in such a scenario. When adopting relatively large sub-word vocabularies we observe increases in performance thanks to transfer learning from a parent M-NMT model, both when translating to and from the under-resourced language. Our proposed approach involving dynamic vocabularies is both practical and effective. We report results on two under-resourced language pairs, i.e. Icelandic-English and Irish-English.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.450}
}

@InProceedings{dobreva-zhou-bawden:2020:LREC,
  author   = {Dobreva, Radina and Zhou, Jie and Bawden, Rachel},
  title   = {Document Sub-structure in Neural Machine Translation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3657--3667},
  abstract = {Current approaches to machine translation (MT) either translate sentences in isolation, disregarding the context they appear in, or model context at the level of the full document, without a notion of any internal structure the document may have. In this work we consider the fact that documents are rarely homogeneous blocks of text, but rather consist of parts covering different topics. Some documents, such as biographies and encyclopedia

entries, have highly predictable, regular structures in which sections are characterised by different topics. We draw inspiration from Louis and Webber (2014) who use this information to improve statistical MT and transfer their proposal into the framework of neural MT. We compare two different methods of including information about the topic of the section within which each sentence is found: one using side constraints and the other using a cache-based model. We create and release the data on which we run our experiments - parallel corpora for three language pairs (Chinese-English, French-English, Bulgarian-English) from Wikipedia biographies, which we extract automatically, preserving the boundaries of sections within the articles.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.451}
}

@InProceedings{raganato-scherrer-tiedemann:2020:LREC,
  author   = {Raganato, Alessandro and Scherrer, Yves and Tiedemann, Jörg},
  title   = {An Evaluation Benchmark for Testing the Word Sense Disambiguation Capabilities of Machine Translation Systems},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3668--3675},
  abstract = {Lexical ambiguity is one of the many challenging linguistic phenomena involved in translation, i.e., translating an ambiguous word with its correct sense. In this respect, previous work has shown that the translation quality of neural machine translation systems can be improved by explicitly modeling the senses of ambiguous words. Recently, several evaluation test sets have been proposed to measure the word sense disambiguation (WSD) capability of machine translation systems. However, to date, these evaluation test sets do not include any training data that would provide a fair setup measuring the sense distributions present within the training data itself. In this paper, we present an evaluation benchmark on WSD for machine translation for 10 language pairs, comprising training data with known sense distributions. Our approach for the construction of the benchmark builds upon the wide-coverage multilingual sense inventory of BabelNet, the multilingual neural parsing pipeline TurkuNLP, and the OPUS collection of translated texts from the web. The test suite is available at http://github.com/Helsinki-NLP/MuCoW.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.452}
}

@InProceedings{nvol-jimenoyepes-neves:2020:LREC,
  author   = {Névéol, Aurélie and Jimeno Yepes, Antonio and Neves, Mariana},
  title   = {MEDLINE as a Parallel Corpus: a Survey to Gain Insight on French-, Spanish- and Portuguese-speaking Authors' Abstract Writing Practice},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3676--3682},
  abstract  = {Background: Parallel corpora are used to train and
evaluate machine translation systems. To alleviate the cost of
producing parallel resources for evaluation campaigns, existing
corpora are leveraged. However, little information may be available
about the methods used for producing the corpus, including
translation direction. Objective: To gain insight on MEDLINE
parallel corpus used in the biomedical task at the Workshop on
Machine Translation in 2019 (WMT 2019). Material and Methods:
Contact information for the authors of MEDLINE articles included in
the English/Spanish (EN/ES), English/French (EN/FR), and English/
Portuguese (EN/PT) WMT 2019 test sets was obtained from PubMed and
publisher websites. The authors were asked about their abstract
writing practices in a survey. Results: The response rate was above
20\%. Authors reported that they are mainly native speakers of
languages other than English. Although manual translation, sometimes
via professional translation services, was commonly used for
abstract translation, authors of articles in the EN/ES and EN/PT
sets also relied on post-edited machine translation. Discussion:
This study provides a characterization of MEDLINE authors' language
skills and abstract writing practices. Conclusion: The information
collected in this study will be used to inform test set design for
the next WMT biomedical task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.453}
}

@InProceedings{mao-EtAl:2020:LREC,
  author    = {Mao, Zhuoyuan  and  Cromieres, Fabien  and  Dabre,
Raj  and  Song, Haiyue  and  Kurohashi, Sadao},
  title     = {JASS: Japanese-specific Sequence to Sequence Pre-
training for Neural Machine Translation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3683--3691},
  abstract  = {Neural machine translation (NMT) needs large parallel
corpora for state-of-the-art translation quality. Low-resource NMT
is typically addressed by transfer learning which leverages large
monolingual or parallel corpora for pre-training. Monolingual pre-
training approaches such as MASS (MAsked Sequence to Sequence) are
extremely effective in boosting NMT quality for languages with small
parallel corpora. However, they do not account for linguistic
information obtained using syntactic analyzers which is known to be
invaluable for several Natural Language Processing (NLP) tasks. To
this end, we propose JASS, Japanese-specific Sequence to Sequence,
```

as a novel pre-training alternative to MASS for NMT involving Japanese as the source or target language. JASS is joint BMASS (Bunsetsu MASS) and BRSS (Bunsetsu Reordering Sequence to Sequence) pre-training which focuses on Japanese linguistic units called bunsetsus. In our experiments on ASPEC Japanese–English and News Commentary Japanese–Russian translation we show that JASS can give results that are competitive with if not better than those given by MASS. Furthermore, we show for the first time that joint MASS and JASS pre-training gives results that significantly surpass the individual methods indicating their complementary nature. We will release our code, pre-trained models and bunsetsu annotated data as resources for researchers to use in their own NLP tasks.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.454}
}

@InProceedings{ive-EtAl:2020:LREC,
  author   = {Ive, Julia and Specia, Lucia and Szoc, Sara and Vanallemeersch, Tom and Van den Bogaert, Joachim and Farah, Eduardo and Maroti, Christine and Ventura, Artur and Khalilov, Maxim},
  title   = {A Post-Editing Dataset in the Legal Domain: Do we Underestimate Neural Machine Translation Quality?},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3692--3697},
  abstract = {We introduce a machine translation dataset for three pairs of languages in the legal domain with post-edited high-quality neural machine translation and independent human references. The data was collected as part of the EU APE-QUEST project and comprises crawled content from EU websites with translation from English into three European languages: Dutch, French and Portuguese. Altogether, the data consists of around 31K tuples including a source sentence, the respective machine translation by a neural machine translation system, a post-edited version of such translation by a professional translator, and - where available - the original reference translation crawled from parallel language websites. We describe the data collection process, provide an analysis of the resulting post-edits and benchmark the data using state-of-the-art quality estimation and automatic post-editing models. One interesting by-product of our post-editing analysis suggests that neural systems built with publicly available general domain data can provide high-quality translations, even though comparison to human references suggests that this quality is quite low. This makes our dataset a suitable candidate to test evaluation metrics. The data is freely available as an ELRC-SHARE resource.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.455}
}

@InProceedings{goyal-mishra-sharma:2020:LREC,
  author   = {Goyal, Vikrant and Mishra, Pruthwik and Sharma,

```
Dipti Misra},
  title     = {Linguistically Informed Hindi-English Neural Machine
Translation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3698--3703},
  abstract  = {Hindi-English Machine Translation is a challenging
problem, owing to multiple factors including the morphological
complexity and relatively free word order of Hindi, in addition to
the lack of sufficient parallel training data. Neural Machine
Translation (NMT) is a rapidly advancing MT paradigm and has shown
promising results for many language pairs, especially in large
training data scenarios. To overcome the data sparsity issue caused
by the lack of large parallel corpora for Hindi-English, we propose
a method to employ additional linguistic knowledge which is encoded
by different phenomena depicted by Hindi. We generalize the
embedding layer of the state-of-the-art Transformer model to
incorporate linguistic features like POS tag, lemma and morph
features to improve the translation performance. We compare the
results obtained on incorporating this knowledge with the baseline
systems and demonstrate significant performance improvements.
Although, the Transformer NMT models have a strong efficacy to learn
language constructs, we show that the usage of specific features
further help in improving the translation performance.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.456}
}

@InProceedings{nagata-morishita:2020:LREC,
  author    = {Nagata, Masaaki  and  Morishita, Makoto},
  title     = {A Test Set for Discourse Translation from Japanese to
English},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3704--3709},
  abstract  = {We made a test set for Japanese-to-English discourse
translation to evaluate the power of context-aware machine
translation. For each discourse phenomenon, we systematically
collected examples where the translation of the second sentence
depends on the first sentence. Compared with a previous study on
test sets for English-to-French discourse translation
\cite{Bawden\_elal\_NAACL2018}, we needed different approaches to
make the data because Japanese has zero pronouns and represents
different senses in different characters. We improved the
translation accuracy using context-aware neural machine translation,
and the improvement mainly reflects the betterment of the
translation of zero pronouns.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.457}
}

@InProceedings{mueller-EtAl:2020:LREC,
  author    = {Mueller, Aaron  and  Nicolai, Garrett  and  McCarthy,
Arya D.  and  Lewis, Dylan  and  Wu, Winston  and  Yarowsky, David},
  title     = {An Analysis of Massively Multilingual Neural Machine
Translation for Low-Resource Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3710--3718},
  abstract  = {In this work, we explore massively multilingual low-
resource neural machine translation. Using translations of the Bible
(which have parallel structure across languages), we train models
with up to 1,107 source languages. We create various multilingual
corpora, varying the number and relatedness of source languages.
Using these, we investigate the best ways to use this many-way
aligned resource for multilingual machine translation. Our
experiments employ a grammatically and phylogenetically diverse set
of source languages during testing for more representative
evaluations. We find that best practices in this domain are highly
language-specific: adding more languages to a training set is often
better, but too many harms performance---the best number depends on
the source language. Furthermore, training on related languages can
improve or degrade performance, depending on the language. As there
is no one-size-fits-most answer, we find that it is critical to
tailor one's approach to the source language and its typology.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.458}
}

@InProceedings{doi-oda-nakazawa:2020:LREC,
  author    = {Doi, Nobushige  and  Oda, Yusuke  and  Nakazawa,
Toshiaki},
  title     = {TDDC: Timely Disclosure Documents Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3719--3726},
  abstract  = {In this paper, we describe the details of the Timely
Disclosure Documents Corpus (TDDC). TDDC was prepared by manually
aligning the sentences from past Japanese and English timely
disclosure documents in PDF format published by companies listed on
the Tokyo Stock Exchange. TDDC consists of approximately 1.4 million
parallel sentences in Japanese and English. TDDC was used as the
official dataset for the 6th Workshop on Asian Translation to
encourage the development of machine translation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.459}
```

}

@InProceedings{morgadodacosta-EtAl:2020:LREC,
  author   = {Morgado da Costa, Luís and V P Winder, Roger and Li, Shu Yun and Lin Tzer Liang, Benedict Christopher and Mackinnon, Joseph and Bond, Francis},
  title   = {Automated Writing Support Using Deep Linguistic Parsers},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {369--377},
  abstract = {This paper introduces a new web system that integrates English Grammatical Error Detection (GED) and course-specific stylistic guidelines to automatically review and provide feedback on student assignments. The system is being developed as a pedagogical tool for English Scientific Writing. It uses both general NLP methods and high precision parsers to check student assignments before they are submitted for grading. Instead of generalized error detection, our system aims to identify, with high precision, specific classes of problems that are known to be common among engineering students. Rather than correct the errors, our system generates constructive feedback to help students identify and correct them on their own. A preliminary evaluation of the system's in-class performance has shown measurable improvements in the quality of student assignments.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.46}
}

@InProceedings{karakanta-negri-turchi:2020:LREC,
  author   = {Karakanta, Alina and Negri, Matteo and Turchi, Marco},
  title   = {MuST-Cinema: a Speech-to-Subtitles corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3727--3734},
  abstract = {Growing needs in localising audiovisual content in multiple languages through subtitles call for the development of automatic solutions for human subtitling. Neural Machine Translation (NMT) can contribute to the automatisation of subtitling, facilitating the work of human subtitlers and reducing turn-around times and related costs. NMT requires high-quality, large, task-specific training data. The existing subtitling corpora, however, are missing both alignments to the source language audio and important information about subtitle breaks. This poses a significant limitation for developing efficient automatic approaches for subtitling, since the length and form of a subtitle directly

depends on the duration of the utterance. In this work, we present MuST-Cinema, a multilingual speech translation corpus built from TED subtitles. The corpus is comprised of (audio, transcription, translation) triplets. Subtitle breaks are preserved by inserting special symbols. We show that the corpus can be used to build models that efficiently segment sentences into subtitles and propose a method for annotating existing subtitling corpora with subtitle breaks, conforming to the constraint of length.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.460}
}

@InProceedings{castilho-popovi-way:2020:LREC,
  author   = {Castilho, Sheila and Popović, Maja and Way, Andy},
  title   = {On Context Span Needed for Machine Translation Evaluation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3735--3742},
  abstract = {Despite increasing efforts to improve evaluation of machine translation (MT) by going beyond the sentence level to the document level, the definition of what exactly constitutes a "document level" is still not clear. This work deals with the context span necessary for a more reliable MT evaluation. We report results from a series of surveys involving three domains and 18 target languages designed to identify the necessary context span as well as issues related to it. Our findings indicate that, despite the fact that some issues and spans are strongly dependent on domain and on the target language, a number of common patterns can be observed so that general guidelines for context-aware MT evaluation can be drawn.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.461}
}

@InProceedings{siripragrada-EtAl:2020:LREC,
  author   = {Siripragrada, Shashank and Philip, Jerin and Namboodiri, Vinay P. and Jawahar, C V},
  title   = {A Multilingual Parallel Corpora Collection Effort for Indian Languages},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3743--3751},
  abstract = {We present sentence aligned parallel corpora across 10 Indian Languages - Hindi, Telugu, Tamil, Malayalam, Gujarati, Urdu, Bengali, Oriya, Marathi, Punjabi, and English - many of which are categorized as low resource. The corpora are compiled from

```
online sources which have content shared across languages. The
corpora presented significantly extends present resources that are
either not large enough or are restricted to a specific domain (such
as health). We also provide a separate test corpus compiled from an
independent online source that can be independently used for
validating the performance in 10 Indian languages. Alongside, we
report on the methods of constructing such corpora using tools
enabled by recent advances in machine translation and cross-lingual
retrieval using deep neural network based methods.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.462}
}

@InProceedings{etchegoyhen-gete:2020:LREC1,
  author    = {Etchegoyhen, Thierry  and  Gete, Harritxu},
  title     = {To Case or not to case: Evaluating Casing Methods for
Neural Machine Translation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3752--3760},
  abstract  = {We present a comparative evaluation of casing methods
for Neural Machine Translation, to help establish an optimal pre-
and post-processing methodology. We trained and compared system
variants on data prepared with the main casing methods available,
namely translation of raw data without case normalisation,
lowercasing with recasing, truecasing, case factors and inline
casing. Machine translation models were prepared on WMT 2017
English-German and English-Turkish datasets, for all translation
directions, and the evaluation includes reference metric results as
well as a targeted analysis of case preservation accuracy. Inline
casing, where case information is marked along lowercased words in
the training data, proved to be the optimal approach overall in
these experiments.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.463}
}

@InProceedings{vradi-EtAl:2020:LREC,
  author    = {Váradi, Tamás  and  Koeva, Svetla  and  Yamalov,
Martin  and  Tadić, Marko  and  Sass, Bálint  and  Nitoń, Bartłomiej
and  Ogrodniczuk, Maciej  and  Pęzik, Piotr  and  Barbu Mititelu,
Verginica  and  Ion, Radu  and  Irimia, Elena  and  Mitrofan, Maria
and  Păiș, Vasile  and  Tufiș, Dan  and  Garabík, Radovan  and
Krek, Simon  and  Repar, Andraz  and  Rihtar, Matjaž  and  Brank,
Janez},
  title     = {The MARCELL Legislative Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {3761--3768},
  abstract  = {This article presents the current outcomes of the
MARCELL CEF Telecom project aiming to collect and deeply annotate a
large comparable corpus of legal documents. The MARCELL corpus
includes 7 monolingual sub-corpora (Bulgarian, Croatian, Hungarian,
Polish, Romanian, Slovak and Slovenian) containing the total body of
respective national legislative documents. These sub-corpora are
automatically sentence split, tokenized, lemmatized and
morphologically and syntactically annotated. The monolingual sub-
corpora are complemented by a thematically related parallel corpus
(Croatian-English). The metadata and the annotations are uniformly
provided for each language specific sub-corpus. Besides the standard
morphosyntactic analysis plus named entity and dependency
annotation, the corpus is enriched with the IATE and EUROVOC labels.
The file format is CoNLL-U Plus Format, containing the ten columns
specific to the CoNLL-U format and four extra columns specific to
our corpora. The MARCELL corpora represents a rich and valuable
source for further studies and developments in machine learning,
cross-lingual terminological data extraction and classification.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.464}
}

@InProceedings{soares-EtAl:2020:LREC,
  author    = {Soares, Felipe  and  Stevenson, Mark  and  Bartolome,
Diego  and  Zaretskaya, Anna},
  title     = {ParaPat: The Multi-Million Sentences Parallel Corpus
of Patents Abstracts},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3769--3774},
  abstract  = {The Google Patents is one of the main important
sources of patents information. A striking characteristic is that
many of its abstracts are presented in more than one language, thus
making it a potential source of parallel corpora. This article
presents the development of a parallel corpus from the open access
Google Patents dataset in 74 language pairs, comprising more than 68
million sentences and 800 million tokens. Sentences were
automatically aligned using the Hunalign algorithm for the largest
22 language pairs, while the others were abstract (i.e. paragraph)
aligned. We demonstrate the capabilities of our corpus by training
Neural Machine Translation (NMT) models for the main 9 language
pairs, with a total of 18 models. Our parallel corpus is freely
available in TSV format and with a SQLite database, with
complementary information regarding patent metadata.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.465}
}

@InProceedings{liu-zhang:2020:LREC,
  author    = {Liu, Siyou  and  Zhang, Xiaojun},
  title     = {Corpora for Document-Level Neural Machine
```

```
Translation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3775--3781},
  abstract  = {Instead of translating sentences in isolation,
document-level machine translation aims to capture discourse
dependencies across sentences by considering a document as a whole.
In recent years, there have been more interests in modelling larger
context for the state-of-the-art neural machine translation (NMT).
Although various document-level NMT models have shown significant
improvements, there nonetheless exist three main problems: 1)
compared with sentence-level translation tasks, the data for
training robust document-level models are relatively low-resourced;
2) experiments in previous work are conducted on their own datasets
which vary in size, domain and language; 3) proposed approaches are
implemented on distinct NMT architectures such as recurrent neural
networks (RNNs) and self-attention networks (SANs). In this paper,
we aims to alleviate the low-resource and under-universality
problems for document-level NMT. First, we collect a large number of
existing document-level corpora, which covers 7 language pairs and 6
domains. In order to address resource sparsity, we construct a novel
document parallel corpus in Chinese-Portuguese, which is a non-
English-centred and low-resourced language pair. Besides, we
implement and evaluate the commonly-cited document-level method on
top of the advanced Transformer model with universal settings.
Finally, we not only demonstrate the effectiveness and universality
of document-level NMT, but also release the preprocessed data,
source code and trained models for comparison and reproducibility.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.466}
}

@InProceedings{aulamo-EtAl:2020:LREC,
  author    = {Aulamo, Mikko  and  Sulubacak, Umut  and  Virpioja,
Sami  and  Tiedemann, Jörg},
  title     = {OpusTools and Parallel Corpus Diagnostics},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3782--3789},
  abstract  = {This paper introduces OpusTools, a package for
downloading and processing parallel corpora included in the OPUS
corpus collection. The package implements tools for accessing
compressed data in their archived release format and make it
possible to easily convert between common formats. OpusTools also
includes tools for language identification and data filtering as
well as tools for importing data from various sources into the OPUS
format. We show the use of these tools in parallel corpus creation
```

and data diagnostics. The latter is especially useful for the identification of potential problems and errors in the extensive data set. Using these tools, we can now monitor the validity of data sets and improve the overall quality and consitency of the data collection.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.467}
}

@InProceedings{fonteyne-tezcan-macken:2020:LREC,
  author   = {Fonteyne, Margot and Tezcan, Arda and Macken, Lieve},
  title   = {Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3790--3798},
  abstract = {Several studies (covering many language pairs and translation tasks) have demonstrated that translation quality has improved enormously since the emergence of neural machine translation systems. This raises the question whether such systems are able to produce high-quality translations for more creative text types such as literature and whether they are able to generate coherent translations on document level. Our study aimed to investigate these two questions by carrying out a document-level evaluation of the raw NMT output of an entire novel. We translated Agatha Christie's novel The Mysterious Affair at Styles with Google’s NMT system from English into Dutch and annotated it in two steps: first all fluency errors, then all accuracy errors. We report on the overall quality, determine the remaining issues, compare the most frequent error types to those in general-domain MT, and investigate whether any accuracy and fluency errors co-occur regularly. Additionally, we assess the inter-annotator agreement on the first chapter of the novel.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.468}
}

@InProceedings{etchegoyhen-gete:2020:LREC2,
  author   = {Etchegoyhen, Thierry and Gete, Harritxu},
  title   = {Handle with Care: A Case Study in Comparable Corpora Exploitation for Neural Machine Translation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3799--3807},
  abstract = {We present the results of a case study in the exploitation of comparable corpora for Neural Machine Translation. A

```
large comparable corpus for Basque-Spanish was prepared, on the
basis of independently-produced news by the Basque public
broadcaster EiTB, and we discuss the impact of various techniques to
exploit the original data in order to determine optimal variants of
the corpus. In particular, we show that filtering in terms of
alignment thresholds and length-difference outliers has a
significant impact on translation quality. The impact of tags
identifying comparable data in the training datasets is also
evaluated, with results indicating that this technique might be
useful to help the models discriminate noisy information, in the
form of informational imbalance between aligned sentences. The final
corpus was prepared according to the experimental results and is
made available to the scientific community for research purposes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.469}
}

@InProceedings{gretter-EtAl:2020:LREC,
  author    = {Gretter, Roberto  and  Matassoni, Marco  and  Bannò,
Stefano  and  Daniele, Falavigna},
  title     = {TLT-school: a Corpus of Non Native Children Speech},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {378--385},
  abstract  = {This paper describes ``TLT-school'' a corpus of
speech utterances collected in schools of northern Italy for
assessing the performance of students learning both English and
German. The corpus was recorded in the years 2017 and 2018 from
students aged between nine and sixteen years, attending primary,
middle and high school. All utterances have been scored, in terms of
some predefined proficiency indicators, by human experts. In
addition, most of utterances recorded in 2017 have been manually
transcribed carefully. Guidelines and procedures used for manual
transcriptions of utterances will be described in detail, as well as
results achieved by means of an automatic speech recognition system
developed by us. Part of the corpus is going to be freely
distributed to scientific community particularly interested both in
non-native speech recognition and automatic assessment of second
language proficiency.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.47}
}

@InProceedings{tiedemann-EtAl:2020:LREC,
  author    = {Tiedemann, Jörg  and  Nieminen, Tommi  and  Aulamo,
Mikko  and  Kanerva, Jenna  and  Leino, Akseli  and  Ginter, Filip
and  Papula, Niko},
  title     = {The FISKMÖ Project: Resources and Tools for Finnish-
Swedish Machine Translation and Cross-Linguistic Research},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {3808--3815},
  abstract  = {This paper presents FISKMÖ, a project that focuses on
the development of resources and tools for cross-linguistic research
and machine translation between Finnish and Swedish. The goal of the
project is the compilation of a massive parallel corpus out of
translated material collected from web sources, public and private
organisations and language service providers in Finland with its two
official languages. The project also aims at the development of open
and freely accessible translation services for those two languages
for the general purpose and for domain-specific use. We have
released new data sets with over 3 million translation units, a
benchmark test set for MT development, pre-trained neural MT models
with high coverage and competitive performance and a self-contained
MT plugin for a popular CAT tool. The latter enables offline
translation without dependencies on external services making it
possible to work with highly sensitive data without compromising
security concerns.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.470}
}

@InProceedings{zaninello-birch:2020:LREC,
  author    = {Zaninello, Andrea  and  Birch, Alexandra},
  title     = {Multiword Expression aware Neural Machine
Translation},
  booktitle     = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month         = {May},
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {3816--3825},
  abstract  = {Multiword Expressions (MWEs) are a frequently
occurring phenomenon found in all natural languages that is of great
importance to linguistic theory, natural language processing
applications, and machine translation systems. Neural Machine
Translation (NMT) architectures do not handle these expressions well
and previous studies have rarely addressed MWEs in this framework.
In this work, we show that annotation and data augmentation, using
external linguistic resources, can improve both translation of MWEs
that occur in the source, and the generation of MWEs on the target,
and increase performance by up to 5.09 BLEU points on MWE test sets.
We also devise a MWE score to specifically assess the quality of MWE
translation which agrees with human evaluation. We make available
the MWE score implementation – along with MWE-annotated training
sets and corpus-based lists of MWEs – for reproduction and
extension.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.471}
}

@InProceedings{kim-colineau:2020:LREC,
  author    = {Kim, Myung Hee  and  Colineau, Nathalie},
```

```
  title     = {An Enhanced Mapping Scheme of the Universal Part-Of-
Speech for Korean},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3826--3833},
  abstract  = {When mapping a language specific Part-Of-Speech (POS)
tag set to the Universal POS tag set (UPOS), it is critical to
consider the individual language’s linguistic features and the UPOS
definitions. In this paper, we present an enhanced Sejong POS
mapping to the UPOS in accordance with the Korean linguistic
typology and the substantive definitions of the UPOS categories.
This work updated one third of the Sejong POS mapping to the UPOS.
We also introduced a new mapping for the KAIST POS tag set, another
widely used Korean POS tag set, to the UPOS.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.472}
}

@InProceedings{alkhairy-jafri-smith:2020:LREC,
  author    = {Alkhairy, Maha  and  Jafri, Afshan  and  Smith,
David},
  title     = {Finite State Machine Pattern-Root Arabic
Morphological Generator, Analyzer and Diacritizer},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3834--3841},
  abstract  = {We describe and evaluate the Finite-State Arabic
Morphologizer (FSAM) – a concatenative (prefix-stem-suffix) and
templatic (root- pattern) morphologizer that generates and analyzes
undiacritized Modern Standard Arabic (MSA) words, and diacritizes
them. Our bidirectional unified-architecture finite state machine
(FSM) is based on morphotactic MSA grammatical rules. The FSM models
the root-pattern structure related to semantics and syntax, making
it readily scalable unlike stem-tabulations in prevailing systems.
We evaluate the coverage and accuracy of our model, with coverage
being percentage of words in Tashkeela (a large corpus) that can be
analyzed. Accuracy is computed against a gold standard, comprising
words and properties, created from the intersection of UD PADT
treebank and Tashkeela. Coverage of analysis (extraction of root and
properties from word) is 82\%. Accuracy results are: root computed
from a word (92\%), word generation from a root (100\%), non-root
properties of a word (97\%), and diacritization (84\%). FSAM’s non-
root results match or surpass MADAMIRA’s, and root result
comparisons are not made because of the concatenative nature of
publicly available morphologizers.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.473}
}
```

```
@InProceedings{keleg-EtAl:2020:LREC,
  author    = {Keleg, Amr  and  Tyers, Francis  and  Howell, Nick
and  Pirinen, Tommi},
  title     = {An Unsupervised Method for Weighting Finite-state
Morphological Analyzers},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3842--3850},
  abstract  = {Morphological analysis is one of the tasks that have
been studied for years. Different techniques have been used to
develop models for performing morphological analysis. Models based
on finite state transducers have proved to be more suitable for
languages with low available resources. In this paper, we have
developed a method for weighting a morphological analyzer built
using finite state transducers in order to disambiguate its results.
The method is based on a word2vec model that is trained in a
completely unsupervised way using raw untagged corpora and is able
to capture the semantic meaning of the words. Most of the methods
used for disambiguating the results of a morphological analyzer
relied on having tagged corpora that need to manually built.
Additionally, the method developed uses information about the token
irrespective of its context unlike most of the other techniques that
heavily rely on the word's context to disambiguate its set of
candidate analyses.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.474}
}

@InProceedings{bollegala-EtAl:2020:LREC,
  author    = {Bollegala, Danushka  and  Kiryo, Ryuichi  and
Tsujino, Kosuke  and  Yukawa, Haruki},
  title     = {Language-Independent Tokenisation Rivals Language-
Specific Tokenisation for Word Similarity Prediction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3851--3860},
  abstract  = {Language-independent tokenisation (LIT) methods that
do not require labelled language resources or lexicons have recently
gained popularity because of their applicability in resource-poor
languages. Moreover, they compactly represent a language using a
fixed size vocabulary and can efficiently handle unseen or rare
words. On the other hand, language-specific tokenisation (LST)
methods have a long and established history, and are developed using
carefully created lexicons and training resources. Unlike subtokens
produced by LIT methods, LST methods produce valid morphological
subwords. Despite the contrasting trade-offs between LIT vs. LST
```

```
methods, their performance on downstream NLP tasks remain unclear. In this paper, we empirically compare the two approaches using semantic similarity measurement as an evaluation task across a diverse set of languages. Our experimental results covering eight languages show that LST consistently outperforms LIT when the vocabulary size is large, but LIT can produce comparable or better results than LST in many languages with comparatively smaller (i.e. less than 100K words) vocabulary sizes, encouraging the use of LIT when language-specific resources are unavailable, incomplete or a smaller model is required. Moreover, we find that smoothed inverse frequency (SIF) to be an accurate method to create word embeddings from subword embeddings for multilingual semantic similarity prediction tasks. Further analysis of the nearest neighbours of tokens show that semantically and syntactically related tokens are closely embedded in subword embedding spaces.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.475}
}

@InProceedings{nikiforos-kermanidis:2020:LREC,
  author    = {Nikiforos, Maria Nefeli  and  Kermanidis, Katia Lida},
  title     = {A Supervised Part-Of-Speech Tagger for the Greek Language of the Social Web},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3861--3867},
  abstract  = {The increasing volume of communication via microblogging messages on social networks has created the need for efficient Natural Language Processing (NLP) tools, especially for unstructured text processing. Extracting information from unstructured social text is one of the most demanding NLP tasks. This paper presents the first part-of-speech tagged data set of social text in Greek, as well as the first supervised part-of-speech tagger developed for such data sets.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.476}
}

@InProceedings{jonker-deruijt-degruijl:2020:LREC,
  author    = {Jonker, Anne  and  de Ruijt, Corné  and  de Gruijl, Jornt},
  title     = {Bag \& Tag'em - A New Dutch Stemmer},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3868--3876},
  abstract  = {We propose a novel stemming algorithm that is both robust and accurate compared to state-of-the-art solutions, yet
```

```
addresses several of the problems that current stemmers face in the Dutch language. The main issue is that most current stemmers cannot handle 3rd person singular forms of verbs and many irregular words and conjugations, unless a (nearly) brute-force approach is used. Our algorithm combines a new tagging module with a stemmer that uses tag-specific sets of rigid rules: the Bag \& Tag'em (BT) algorithm. The tagging module is developed and evaluated using three algorithms: Multinomial Logistic Regression (MLR), Neural Network (NN) and Extreme Gradient Boosting (XGB). The stemming module's performance is compared with that of current state-of-the-art stemming algorithms for the Dutch Language. Even though there is still room for improvement, the new BT algorithm performs well in the sense that it is more accurate than the current stemmers and faster than brute-force-like algorithms. The code and data used for this paper can be found at: https://github.com/Anne-Jonker/Bag-Tag-em.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.477}
}

@InProceedings{hathout-EtAl:2020:LREC,
  author    = {Hathout, Nabil  and  Sajous, Franck  and  Calderone, Basilio  and  Namer, Fiammetta},
  title     = {Glawinette: a Linguistically Motivated Derivational Description of French Acquired from GLAWI},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3877--3885},
  abstract  = {Glawinette is a derivational lexicon of French that will be used to feed the Démonette database. It has been created from the GLAWI machine readable dictionary. We collected couples of words from the definitions and the morphological sections of the dictionary and then selected the ones that form regular formal analogies and that instantiate frequent enough formal patterns. The graph structure of the morphological families has then been used to identify for each couple of lexemes derivational patterns that are close to the intuition of the morphologists.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.478}
}

@InProceedings{sahala-EtAl:2020:LREC2,
  author    = {Sahala, Aleksi  and  Silfverberg, Miikka  and  Arppe, Antti  and  Lindén, Krister},
  title     = {BabyFST - Towards a Finite-State Based Computational Model of Ancient Babylonian},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {3886--3894},
  abstract  = {Akkadian is a fairly well resourced extinct language
that does not yet have a comprehensive morphological analyzer
available. In this paper we describe a general finite-state based
morphological model for Babylonian, a southern dialect of the
Akkadian language, that can achieve a coverage up to 97.3\% and
recall up to 93.7\% on lemmatization and POS-tagging task on token
level from a transcribed input. Since Akkadian word forms exhibit a
high degree of morphological ambiguity, in that only 20.1\% of
running word tokens receive a single unambiguous analysis, we
attempt a first pass at weighting our finite-state transducer, using
existing extensive Akkadian corpora which have been partially
validated for their lemmas and parts-of-speech but not the entire
morphological analyses. The resultant weighted finite-state
transducer yields a moderate improvement so that for 57.4\% of the
word tokens the highest ranked analysis is the correct one. We
conclude with a short discussion on how morphological ambiguity in
the analysis of Akkadian could be further reduced with improvements
in the training data used in weighting the finite-state transducer
as well as through other, context-based techniques.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.479}
}

@InProceedings{katinskaia-ivanova-yangarber:2020:LREC,
  author    = {Katinskaia, Anisia  and  Ivanova, Sardana  and
Yangarber, Roman},
  title     = {Toward a Paradigm Shift in Collection of Learner
Corpora},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {386--391},
  abstract  = {We present the first version of the longitudinal
Revita Learner Corpus (ReLCo), for Russian. In contrast to
traditional learner corpora, ReLCo is collected and annotated fully
automatically, while students perform exercises using the Revita
language-learning platform. The corpus currently contains 8 422
sentences exhibiting several types of errors—grammatical, lexical,
orthographic, etc.—which were committed by learners during practice
and were automatically annotated by Revita. The corpus provides
valuable information about patterns of learner errors and can be
used as a language resource for a number of research tasks, while
its creation is much cheaper and faster than for traditional learner
corpora. A crucial advantage of ReLCo that it grows continually
while learners practice with Revita, which opens the possibility of
creating an unlimited learner resource with longitudinal data
collected over time. We make the pilot version of the Russian ReLCo
publicly available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.48}
}
```

```
@InProceedings{khalifa-zalmout-habash:2020:LREC,
  author    = {Khalifa, Salam  and  Zalmout, Nasser  and  Habash,
Nizar},
  title     = {Morphological Analysis and Disambiguation for Gulf
Arabic: The Interplay between Resources and Methods},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3895--3904},
  abstract  = {In this paper we present the first full morphological
analysis and disambiguation system for Gulf Arabic. We use an
existing state-of-the-art morphological disambiguation system to
investigate the effects of different data sizes and different
combinations of morphological analyzers for Modern Standard Arabic,
Egyptian Arabic, and Gulf Arabic. We find that in very low settings,
morphological analyzers help boost the performance of the full
morphological disambiguation task. However, as the size of resources
increase, the value of the morphological analyzers decreases.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.480}
}

@InProceedings{metheniti-neumann:2020:LREC,
  author    = {Metheniti, Eleni  and  Neumann, Guenter},
  title     = {Wikinflection Corpus: A (Better) Multilingual,
Morpheme-Annotated Inflectional Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3905--3912},
  abstract  = {Multilingual, inflectional corpora are a scarce
resource in the NLP community, especially corpora with annotated
morpheme boundaries. We are evaluating a generated, multilingual
inflectional corpus with morpheme boundaries, generated from the
English Wiktionary (Metheniti and Neumann, 2018), against the
largest, multilingual, high-quality inflectional corpus of the
UniMorph project (Kirov et al., 2018). We confirm that the generated
Wikinflection corpus is not of such quality as UniMorph, but we were
able to extract a significant amount of words from the intersection
of the two corpora. Our Wikinflection corpus benefits from the
morpheme segmentations of Wiktionary/Wikinflection and from the
manually-evaluated morphological feature tags of the UniMorph
project, and has 216K lemmas and 5.4M word forms, in a total of 68
languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.481}
}

@InProceedings{tran-EtAl:2020:LREC,
  author    = {Tran, Oanh  and  Pham, Tu  and  Dang, Vu  and
```

Nguyen, Bang},
  title   = {Introducing a Large-Scale Dataset for Vietnamese POS Tagging on Conversational Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3913--3921},
  abstract = {This paper introduces a large-scale human-labeled dataset for the Vietnamese POS tagging task on conversational texts. To this end, wepropose a new tagging scheme (with 36 POS tags) consisting of exclusive tags for special phenomena of conversational words, developthe annotation guideline and manually annotate 16.310K sentences using this guideline. Based on this corpus, a series of state-of-the-art tagging methods has been conducted to estimate their performances. Experimental results showed that the Conditional Random Fields model using both automatically learnt features from deep neural networks and handcrafted features yielded the best performance. Thismodel achieved 93.36\% in the accuracy score which is 1.6\% and 2.7\% higher than the model using either handcrafted features orautomatically-learnt features, respectively. This result is also a little bit higher than the model of fine-tuning BERT by 0.94\% in theaccuracy score. The performance measured on each POS tag is also very high with >90\% in the F1 score for 20 POS tags and >80\%in the F1 score for 11 POS tags. This work provides the public dataset and preliminary results for follow-up research on this interesting direction.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.482}
}

@InProceedings{mccarthy-EtAl:2020:LREC2,
  author   = {McCarthy, Arya D. and Kirov, Christo and Grella, Matteo and Nidhi, Amrit and Xia, Patrick and Gorman, Kyle and Vylomova, Ekaterina and Mielke, Sabrina J. and Nicolai, Garrett and Silfverberg, Miikka and Arkhangelskiy, Timofey and Krizhanovsky, Nataly and Krizhanovsky, Andrew and Klyachko, Elena and Sorokin, Alexey and Mansfield, John and Ernštreits, Valts and Pinter, Yuval and Jacobs, Cassandra L. and Cotterell, Ryan and Hulden, Mans and Yarowsky, David},
  title   = {UniMorph 3.0: Universal Morphology},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3922--3931},
  abstract = {The Universal Morphology (UniMorph) project is a collaborative effort providing broad-coverage instantiated normalized morphological paradigms for hundreds of diverse world languages. The project comprises two major thrusts: a language-independent feature schema for rich morphological annotation and a

```
type-level resource of annotated data in diverse languages realizing
that schema. We have implemented several improvements to the
extraction pipeline which creates most of our data, so that it is
both more complete and more correct. We have added 66 new languages,
as well as new parts of speech for 12 languages. We have also
amended the schema in several ways. Finally, we present three new
community tools: two to validate data for resource creators, and one
to make morphological data available from the command line. UniMorph
is based at the Center for Language and Speech Processing (CLSP) at
Johns Hopkins University in Baltimore, Maryland. This paper details
advances made to the schema, tooling, and dissemination of project
resources since the UniMorph 2.0 release described at LREC 2018.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.483}
}

@InProceedings{mikeleni-tadi:2020:LREC,
  author    = {Mikelenić, Bojana  and  Tadić, Marko},
  title     = {Building the Spanish-Croatian Parallel Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3932--3936},
  abstract  = {This paper describes the building of the first
Spanish-Croatian unidirectional parallel corpus, which has been
constructed at the Faculty of Humanities and Social Sciences of the
University of Zagreb. The corpus is comprised of eleven Spanish
novels and their translations to Croatian done by six different
professional translators. All the texts were published between 1999
and 2012. The corpus has more than 2 Mw, with approximately 1 Mw for
each language. It was automatically sentence segmented and aligned,
as well as manually post-corrected, and contains 71,778 translation
units. In order to protect the copyright and to make the corpus
available under permissive CC-BY licence, the aligned translation
units are shuffled. This limits the usability of the corpus for
research of language units at sentence and lower language levels
only. There are two versions of the corpus in TMX format that will
be available for download through META-SHARE and CLARIN ERIC
infrastructure. The former contains plain TMX, while the latter is
lemmatised and POS-tagged and stored in the aTMX format.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.484}
}

@InProceedings{vodolazsky:2020:LREC,
  author    = {Vodolazsky, Daniil},
  title     = {DerivBase.Ru: a Derivational Morphology Resource for
Russian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {3937--3943},
  abstract  = {Russian morphology has been studied for decades, but
there is still no large high coverage resource that contains the
derivational families (groups of words that share the same root) of
Russian words. The number of words used in different areas of the
language grows rapidly, thus the human-made dictionaries published
long time ago cannot cover the neologisms and the domain-specific
lexicons. To fill such resource gap, we have developed a rule-based
framework for deriving words and we applied it to build a
derivational morphology resource named DerivBase.Ru, which we
introduce in this paper.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.485}
}

@InProceedings{grnroos-virpioja-kurimo:2020:LREC,
  author    = {Grönroos, Stig-Arne  and  Virpioja, Sami  and
Kurimo, Mikko},
  title     = {Morfessor EM+Prune: Improved Subword Segmentation
with Expectation Maximization and Pruning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3944--3953},
  abstract  = {Data-driven segmentation of words into subword units
has been used in various natural language processing applications
such as automatic speech recognition and statistical machine
translation for almost 20 years. Recently it has became more widely
adopted, as models based on deep neural networks often benefit from
subword units even for morphologically simpler languages. In this
paper, we discuss and compare training algorithms for a unigram
subword model, based on the Expectation Maximization algorithm and
lexicon pruning. Using English, Finnish, North Sami, and Turkish
data sets, we show that this approach is able to find better
solutions to the optimization problem defined by the Morfessor
Baseline model than its original recursive training algorithm. The
improved optimization also leads to higher morphological
segmentation accuracy when compared to a linguistic gold standard.
We publish implementations of the new algorithms in the widely-used
Morfessor software package.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.486}
}

@InProceedings{stankovic-EtAl:2020:LREC,
  author    = {Stankovic, Ranka  and  Šandrih, Branislava  and
Krstev, Cvetana  and  Utvić, Miloš  and  Skoric, Mihailo},
  title     = {Machine Learning and Deep Neural Network-Based
Lemmatization and Morphosyntactic Tagging for Serbian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {3954--3962},
  abstract  = {The training of new tagger models for Serbian is
primarily motivated by the enhancement of the existing tagset with
the grammatical category of a gender. The harmonization of resources
that were manually annotated within different projects over a long
period of time was an important task, enabled by the development of
tools that support partial automation. The supporting tools take
into account different taggers and tagsets. This paper focuses on
TreeTagger and spaCy taggers, and the annotation schema alignment
between Serbian morphological dictionaries, MULTEXT-East and
Universal Part-of-Speech tagset. The trained models will be used to
publish the new version of the Corpus of Contemporary Serbian as
well as the Serbian literary corpus. The performance of developed
taggers were compared and the impact of training set size was
investigated, which resulted in around 98\% PoS-tagging precision
per token for both new models. The sr\_basic annotated dataset will
also be published.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.487}
}

@InProceedings{nicolai-EtAl:2020:LREC,
  author    = {Nicolai, Garrett  and  Lewis, Dylan  and  McCarthy,
Arya D.  and  Mueller, Aaron  and  Wu, Winston  and  Yarowsky,
David},
  title     = {Fine-grained Morphosyntactic Analysis and Generation
Tools for More Than One Thousand Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3963--3972},
  abstract  = {Exploiting the broad translation of the Bible into
the world's languages, we train and distribute morphosyntactic tools
for approximately one thousand languages, vastly outstripping
previous distributions of tools devoted to the processing of
inflectional morphology. Evaluation of the tools on a subset of
available inflectional dictionaries demonstrates strong initial
models, supplemented and improved through ensembling and dictionary-
based reranking. Likewise, a novel type-to-token based evaluation
metric allows us to confirm that models generalize well across rare
and common forms alike},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.488}
}

@InProceedings{balabel-EtAl:2020:LREC,
  author    = {Balabel, Mohamed  and  Hamed, Injy  and  Abdennadher,
Slim  and  Vu, Ngoc Thang  and  Çetinoğlu, Özlem},
  title     = {Cairo Student Code-Switch (CSCS) Corpus: An Annotated
Egyptian Arabic-English Corpus},
```

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3973--3977},
  abstract  = {Code-switching has become a prevalent phenomenon
across many communities. It poses a challenge to NLP researchers,
mainly due to the lack of available data needed for training and
testing applications. In this paper, we introduce a new resource: a
corpus of Egyptian- Arabic code-switch speech data that is fully
tokenized, lemmatized and annotated for part-of-speech tags. Beside
the corpus itself, we provide annotation guidelines to address the
unique challenges of annotating code-switch data. Another challenge
that we address is the fact that Egyptian Arabic orthography and
grammar are not standardized.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.489}
}

@InProceedings{daris-EtAl:2020:LREC1,
  author    = {Darǵis, Roberts  and  Auziņa, Ilze  and  Levāne-
Petrova, Kristīne  and  Kaija, Inga},
  title     = {Quality Focused Approach to a Learner Corpus
Development},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {392--396},
  abstract  = {The paper presents quality focused approach to a
learner corpus development. The methodology was developed with
multiple design considerations put in place to make the annotation
process easier and at the same time reduce the amount of mistakes
that could be introduced due to inconsistent text correction or
carelessness. The approach suggested in this paper consists of
multiple parts: comparison of digitized texts by several annotators,
text correction, automated morphological analysis, and manual review
of annotations. The described approach is used to create Latvian
Language Learner corpus (LaVA) which is part of a currently ongoing
project Development of Learner corpus of Latvian: methods, tools and
applications.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.49}
}

@InProceedings{sorokin:2020:LREC,
  author    = {Sorokin, Alexey},
  title     = {Getting More Data for Low-resource Morphological
Inflection: Language Models and Data Augmentation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {3978--3983},
  abstract  = {We investigate how to improve quality of low-resource
morphological inflection without annotating more data. We examine
two methods, language models and data augmentation. We show that the
model whose decoder that additionally uses the states of the
langauge model improves the model quality by 1.5\% in combination
with both baselines. We also demonstrate that the augmentation of
data improves performance by 9\% in average when adding $1000$
artificially generated word forms to the dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.490}
}

@InProceedings{zen-solak:2020:LREC,
  author    = {Özenç, Berke  and  Solak, Ercan},
  title     = {Visual Modeling of Turkish Morphology},
  booktitle     = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month         = {May},
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {3984--3990},
  abstract  = {In this paper, we describe the steps in a visual
modeling of Turkish morphology using diagramming tools. We aimed to
make modeling easier and more maintainable while automating much of
the code generation. We released the resulting analyzer, MorTur, and
the diagram conversion tool, DiaMor as free, open-source utilities.
MorTur analyzer is also publicly available on its web page as a web
service. MorTur and DiaMor are part of our ongoing efforts in
building a set of natural language processing tools for Turkic
languages under a consistent framework.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.491}
}

@InProceedings{daason-EtAl:2020:LREC,
  author    = {Daðason, Jón  and  Mollberg, David  and  Loftsson,
Hrafn  and  Bjarnadóttir, Kristín},
  title     = {Kvistur 2.0: a BiLSTM Compound Splitter for
Icelandic},
  booktitle     = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month         = {May},
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {3991--3995},
  abstract  = {In this paper, we present a character-based BiLSTM
model for splitting Icelandic compound words, and show how varying
amounts of training data affects the performance of the model.
Compounding is highly productive in Icelandic, and new compounds are
constantly being created. This results in a large number of out-of-
```

```
vocabulary (OOV) words, negatively impacting the performance of many
NLP tools. Our model is trained on a dataset of 2.9 million unique
word forms and their constituent structures from the Database of
Icelandic Morphology. The model learns how to split compound words
into two parts and can be used to derive the constituent structure
of any word form. Knowing the constituent structure of a word form
makes it possible to generate the optimal split for a given task,
e.g., a full split for subword tokenization, or, in the case of
part-of-speech tagging, splitting an OOV word until the largest
known morphological head is found. The model outperforms other
previously published methods when evaluated on a corpus of manually
split word forms. This method has been integrated into Kvistur, an
Icelandic compound word analyzer.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.492}
}

@InProceedings{mott-EtAl:2020:LREC,
  author    = {Mott, Justin  and  Bies, Ann  and  Strassel,
Stephanie  and  Kodner, Jordan  and  Richter, Caitlin  and  Xu,
Hongzhi  and  Marcus, Mitchell},
  title     = {Morphological Segmentation for Low Resource
Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {3996--4002},
  abstract  = {This paper describes a new morphology resource
created by Linguistic Data Consortium and the University of
Pennsylvania for the DARPA LORELEI Program. The data consists of
approximately 2000 tokens annotated for morphological segmentation
in each of 9 low resource languages, along with root information for
7 of the languages. The languages annotated show a broad diversity
of typological features. A minimal annotation scheme for
segmentation was developed such that it could capture the patterns
of a wide range of languages and also be performed reliably by non-
linguist annotators. The basic annotation guidelines were designed
to be language-independent, but included language-specific
morphological paradigms and other specifications. The resulting
annotated corpus is designed to support and stimulate the
development of unsupervised morphological segmenters and analyzers
by providing a gold standard for their evaluation on a more
typologically diverse set of languages than has previously been
available. By providing root annotation, this corpus is also a step
toward supporting research in identifying richer morphological
structures than simple morpheme boundaries.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.493}
}

@InProceedings{wenzek-EtAl:2020:LREC,
  author    = {Wenzek, Guillaume  and  Lachaux, Marie-Anne  and
Conneau, Alexis  and  Chaudhary, Vishrav  and  Guzmán, Francisco
```

```
and  Joulin, Armand  and  Grave, Edouard},
  title     = {CCNet: Extracting High Quality Monolingual Datasets
from Web Crawl Data},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4003--4012},
  abstract  = {Pre-training text representations have led to
significant improvements in many areas of natural language
processing. The quality of these models benefits greatly from the
size of the pretraining corpora as long as its quality is preserved.
In this paper, we describe an automatic pipeline to extract massive
high-quality monolingual datasets from Common Crawl for a variety of
languages. Our pipeline follows the data processing introduced in
fastText (Mikolov et al., 2017; Grave et al., 2018), that
deduplicates documents and identifies their language. We augment
this pipeline with a filtering step to select documents that are
close to high quality corpora like Wikipedia.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.494}
}

@InProceedings{doval-EtAl:2020:LREC,
  author    = {Doval, Yerai  and  Camacho-Collados, Jose  and
Espinosa Anke, Luis  and  Schockaert, Steven},
  title     = {On the Robustness of Unsupervised and Semi-supervised
Cross-lingual Word Embedding Learning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4013--4023},
  abstract  = {Cross-lingual word embeddings are vector
representations of words in different languages where words with
similar meaning are represented by similar vectors, regardless of
the language. Recent developments which construct these embeddings
by aligning monolingual spaces have shown that accurate alignments
can be obtained with little or no supervision, which usually comes
in the form of bilingual dictionaries. However, the focus has been
on a particular controlled scenario for evaluation, and there is no
strong evidence on how current state-of-the-art systems would fare
with noisy text or for language pairs with major linguistic
differences. In this paper we present an extensive evaluation over
multiple cross-lingual embedding models, analyzing their strengths
and limitations with respect to different variables such as target
language, training corpora and amount of supervision. Our
conclusions put in doubt the view that high-quality cross-lingual
embeddings can always be learned without much supervision.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.495}
}
```

```
@InProceedings{zhai-EtAl:2020:LREC,
  author    = {Zhai, Yuming  and  Liu, Lufei  and  Zhong, Xinyi  and
Illouz, Gbariel  and  Vilnat, Anne},
  title     = {Building an English-Chinese Parallel Corpus Annotated
with Sub-sentential Translation Techniques},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4024--4033},
  abstract  = {Human translators often resort to different non-
literal translation techniques besides the literal translation, such
as idiom equivalence, generalization, particularization, semantic
modulation, etc., especially when the source and target languages
have different and distant origins. Translation techniques
constitute an important subject in translation studies, which help
researchers to understand and analyse translated texts. However,
they receive less attention in developing Natural Language
Processing (NLP) applications. To fill this gap, one of our long
term objectives is to have a better semantic control of extracting
paraphrases from bilingual parallel corpora. Based on this goal, we
suggest this hypothesis: it is possible to automatically recognize
different sub-sentential translation techniques. For this original
task, since there is no dedicated data set for English-Chinese, we
manually annotated a parallel corpus of eleven genres. Fifty
sentence pairs for each genre have been annotated in order to
consolidate our annotation guidelines. Based on this data set, we
conducted an experiment to classify between literal and non-literal
translations. The preliminary results confirm our hypothesis. The
corpus and code are available. We hope that this annotated corpus
will be useful for linguistic contrastive studies and for fine-
grained evaluation of NLP tasks, such as automatic word alignment
and machine translation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.496}
}

@InProceedings{nivre-EtAl:2020:LREC,
  author    = {Nivre, Joakim  and  de Marneffe, Marie-Catherine  and
Ginter, Filip  and  Hajic, Jan  and  Manning, Christopher D.  and
Pyysalo, Sampo  and  Schuster, Sebastian  and  Tyers, Francis  and
Zeman, Daniel},
  title     = {Universal Dependencies v2: An Evergrowing
Multilingual Treebank Collection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4034--4043},
  abstract  = {Universal Dependencies is an open community effort to
```

```
create cross-linguistically consistent treebank annotation for many
languages within a dependency-based lexicalist framework. The
annotation consists in a linguistically motivated word segmentation;
a morphological layer comprising lemmas, universal part-of-speech
tags, and standardized morphological features; and a syntactic layer
focusing on syntactic relations between predicates, arguments and
modifiers. In this paper, we describe version 2 of the universal
guidelines (UD v2), discuss the major changes from UD v1 to UD v2,
and give an overview of the currently available treebanks for 90
languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.497}
}

@InProceedings{serratroozen-martnezmartnez:2020:LREC,
  author    = {Serrat Roozen, Iris  and  Martínez Martínez, José
Manuel},
  title     = {EMPAC: an English–Spanish Corpus of Institutional
Subtitles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4044--4053},
  abstract  = {The EuroparlTV Multimedia Parallel Corpus (EMPAC) is
a collection of subtitles in English and Spanish for videos from the
EuropeanParliament’s Multimedia Centre. The corpus has been compiled
with the EMPAC toolkit. The aim of this corpus is to provide a
resource to study institutional subtitling on the one hand, and, on
the other hand, facilitate the analysis of web accessibility to
institutional multimedia content. The corpus covers a time span from
2009 to 2017, it is made up of 4,000 texts amounting to two and half
millions of tokens for every language, corresponding to
approximately 280 hours of video. This paper provides 1) a review of
related corpora; 2) a revision of typical compilation methodologies
of subtitle corpora; 3) a detailed account of the corpus compilation
methodology followed; and, 4) a description of the corpus. In the
conclusion, the key findings are summarised regarding formal aspects
of the subtitles conditioning the accessibility to the multimedia
content of the EuroparlTV.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.498}
}

@InProceedings{kuriyozov-doval-gmezrodrguez:2020:LREC,
  author    = {Kuriyozov, Elmurod  and  Doval, Yerai  and  Gómez-
Rodríguez, Carlos},
  title     = {Cross-Lingual Word Embeddings for Turkic Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

  pages   = {4054--4062},
  abstract = {There has been an increasing interest in learning cross-lingual word embeddings to transfer knowledge obtained from a resource-rich language, such as English, to lower-resource languages for which annotated data is scarce, such as Turkish, Russian, and many others. In this paper, we present the first viability study of established techniques to align monolingual embedding spaces for Turkish, Uzbek, Azeri, Kazakh and Kyrgyz, members of the Turkic family which is heavily affected by the low-resource constraint. Those techniques are known to require little explicit supervision, mainly in the form of bilingual dictionaries, hence being easily adaptable to different domains, including low-resource ones. We obtain new bilingual dictionaries and new word embeddings for these languages and show the steps for obtaining cross-lingual word embeddings using state-of-the-art techniques. Then, we evaluate the results using the bilingual dictionary induction task. Our experiments confirm that the obtained bilingual dictionaries outperform previously-available ones, and that word embeddings from a low-resource language can benefit from resource-rich closely-related languages when they are aligned together. Furthermore, evaluation on an extrinsic task (Sentiment analysis on Uzbek) proves that monolingual word embeddings can, although slightly, benefit from cross-lingual alignments.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.499}
}

@InProceedings{khullar-majmundar-shrivastava:2020:LREC,
  author   = {Khullar, Payal and Majmundar, Kushal and Shrivastava, Manish},
  title   = {NoEl: An Annotated Corpus for Noun Ellipsis in English},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {34--43},
  abstract = {Ellipsis resolution has been identified as an important step to improve the accuracy of mainstream Natural Language Processing (NLP) tasks such as information retrieval, event extraction, dialog systems, etc. Previous computational work on ellipsis resolution has focused on one type of ellipsis, namely Verb Phrase Ellipsis (VPE) and a few other related phenomenon. We extend the study of ellipsis by presenting the No(oun)El(lipsis) corpus - an annotated corpus for noun ellipsis and closely related phenomenon using the first hundred movies of Cornell Movie Dialogs Dataset. The annotations are carried out in a standoff annotation scheme that encodes the position of the licensor, the antecedent boundary, and Part-of-Speech (POS) tags of the licensor and antecedent modifier. Our corpus has 946 instances of exophoric and endophoric noun ellipsis, making it the biggest resource of noun ellipsis in English, to the best of our knowledge. We present a statistical study of our corpus with novel insights on the distribution of noun

```
ellipsis, its licensors and antecedents. Finally, we perform the
tasks of detection and resolution of noun ellipsis with different
classifiers trained on our corpus and report baseline results.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.5}
}

@InProceedings{declercq-vanhoecke:2020:LREC,
  author    = {De Clercq, Orphee  and  Van Hoecke, Senne},
  title     = {An Exploratory Study into Automated Précis Grading},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {397--404},
  abstract  = {Automated writing evaluation is a popular research
field, but the main focus has been on evaluating argumentative
essays. In this paper, we consider a different genre, namely précis
texts. A précis is a written text that provides a coherent summary
of main points of a spoken or written text. We present a corpus of
English précis texts which all received a grade assigned by a
highly-experienced English language teacher and were subsequently
annotated following an exhaustive error typology. With this corpus
we trained a machine learning model which relies on a number of
linguistic, automatic summarization and AWE features. Our results
reveal that this model is able to predict the grade of précis texts
with only a moderate error margin.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.50}
}

@InProceedings{kanayama-iwamoto:2020:LREC,
  author    = {Kanayama, Hiroshi  and  Iwamoto, Ran},
  title     = {How Universal are Universal Dependencies? Exploiting
Syntax for Multilingual Clause-level Sentiment Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4063--4073},
  abstract  = {This paper investigates clause-level sentiment
detection in a multilingual scenario. Aiming at a high-precision,
fine-grained, configurable, and non-biased system for practical use
cases, we have designed a pipeline method that makes the most of
syntactic structures based on Universal Dependencies, avoiding
machine-learning approaches that may cause obstacles to our
purposes. We achieved high precision in sentiment detection for 17
languages and identified the advantages of common syntactic
structures as well as issues stemming from structural differences on
Universal Dependencies. In addition to reusable tips for handling
multilingual syntax, we provide a parallel benchmarking data set for
further research.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.500}
}

@InProceedings{ular-EtAl:2020:LREC,
  author    = {Ulčar, Matej  and  Vaik, Kristiina  and  Lindström,
Jessica  and  Dailidėnaitė, Milda  and  Robnik-Šikonja, Marko},
  title     = {Multilingual Culture-Independent Word Analogy
Datasets},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4074--4080},
  abstract  = {In text processing, deep neural networks mostly use
word embeddings as an input. Embeddings have to ensure that
relations between words are reflected through distances in a high-
dimensional numeric space. To compare the quality of different text
embeddings, typically, we use benchmark datasets. We present a
collection of such datasets for the word analogy task in nine
languages: Croatian, English, Estonian, Finnish, Latvian,
Lithuanian, Russian, Slovenian, and Swedish. We designed the
monolingual analogy task to be much more culturally independent and
also constructed cross-lingual analogy datasets for the involved
languages. We present basic statistics of the created datasets and
their initial evaluation using fastText embeddings.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.501}
}

@InProceedings{costajuss-lilin-espaabonet:2020:LREC,
  author    = {Costa-jussà, Marta R.  and  Li Lin, Pau  and  España-
Bonet, Cristina},
  title     = {GeBioToolkit: Automatic Extraction of Gender-Balanced
Multilingual Corpus of Wikipedia Biographies},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4081--4088},
  abstract  = {We introduce GeBioToolkit, a tool for extracting
multilingual parallel corpora at sentence level, with document and
gender information from Wikipedia biographies. Despite the gender
inequalities present in Wikipedia, the toolkit has been designed to
extract corpus balanced in gender. While our toolkit is customizable
to any number of languages (and different domains), in this work we
present a corpus of 2,000 sentences in English, Spanish and Catalan,
which has been post-edited by native speakers to become a high-
quality dataset for machine translation evaluation. While
GeBioCorpus aims at being one of the first non-synthetic gender-
balanced test datasets, GeBioToolkit aims at paving the path to
standardize procedures to produce gender-balanced datasets.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.502}
}

@InProceedings{johnson-EtAl:2020:LREC,
  author    = {Johnson, Khia A.  and  Babel, Molly  and  Fong, Ivan
and  Yiu, Nancy},
  title     = {SpiCE: A New Open-Access Corpus of Conversational
Bilingual Speech in Cantonese and English},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4089--4095},
  abstract  = {This paper describes the design, collection,
orthographic transcription, and phonetic annotation of SpiCE, a new
corpus of conversational Cantonese-English bilingual speech recorded
in Vancouver, Canada. The corpus includes high-quality recordings of
34 early bilinguals in both English and Cantonese—to date, 27 have
been recorded for a total of 19 hours of participant speech.
Participants completed a sentence reading task, storyboard
narration, and conversational interview in each language.
Transcription and annotation for the corpus are currently underway.
Transcripts produced with Google Cloud Speech-to-Text are available
for all participants, and will be included in the initial SpiCE
corpus release. Hand-corrected orthographic transcripts and force-
aligned phonetic transcripts will be released periodically, and upon
completion for all recordings, comprise the second release of the
corpus. As an open-access language resource, SpiCE will promote
bilingualism research for a typologically distinct pair of
languages, of which Cantonese remains understudied despite there
being millions of speakers around the world. The SpiCE corpus is
especially well-suited for phonetic research on conversational
speech, and enables researchers to study cross-language within-
speaker phenomena for a diverse group of early Cantonese-English
bilinguals. These are areas with few existing high-quality
resources.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.503}
}

@InProceedings{lefever-labat-singh:2020:LREC,
  author    = {Lefever, Els  and  Labat, Sofie  and  Singh,
Pranaydeep},
  title     = {Identifying Cognates in English-Dutch and French-
Dutch by means of Orthographic Information and Cross-lingual Word
Embeddings},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4096--4101},
```

```
  abstract  = {This paper investigates the validity of combining
more traditional orthographic information with cross-lingual word
embeddings to identify cognate pairs in English-Dutch and French-
Dutch. In a first step, lists of potential cognate pairs in English-
Dutch and French-Dutch are manually labelled. The resulting gold
standard is used to train and evaluate a multi-layer perceptron that
can distinguish cognates from non-cognates. Fifteen orthographic
features capture string similarities between source and target
words, while the cosine similarity between their word embeddings
represents the semantic relation between these words. By adding
domain-specific information to pretrained fastText embeddings, we
are able to obtain good embeddings for words that did not yet have a
pretrained embedding (e.g. Dutch compound nouns). These embeddings
are then aligned in a cross-lingual vector space by exploiting their
structural similarity (cf. adversarial learning). Our results
indicate that although the classifier already achieves good results
on the basis of orthographic information, the performance further
improves by including semantic information in the form of cross-
lingual word embeddings.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.504}
}

@InProceedings{kunilovskaya-lapshinovakoltunski:2020:LREC,
  author    = {Kunilovskaya, Maria  and  Lapshinova-Koltunski,
Ekaterina},
  title     = {Lexicogrammatic translationese across two targets and
competence levels},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4102--4112},
  abstract  = {This research employs genre-comparable data from a
number of parallel and comparable corpora to explore the specificity
of translations from English into German and Russian produced by
students and professional translators. We introduce an elaborate set
of human-interpretable lexicogrammatic translationese indicators and
calculate the amount of translationese manifested in the data for
each target language and translation variety. By placing
translations into the same feature space as their sources and the
genre-comparable non-translated reference texts in the target
language, we observe two separate translationese effects: a shift of
translations into the gap between the two languages and a shift away
from either language. These trends are linked to the features that
contribute to each of the effects. Finally, we compare the
translation varieties and find out that the professionalism levels
seem to have some correlation with the amount and types of
translationese detected, while each language pair demonstrates a
specific socio-linguistically determined combination of the
translationese effects.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.505}
}
```

```
@InProceedings{asgari-EtAl:2020:LREC,
  author    = {Asgari, Ehsaneddin  and  Braune, Fabienne  and  Roth, Benjamin  and  Ringlstetter, Christoph  and  Mofrad, Mohammad},
  title     = {UniSent: Universal Adaptable Sentiment Lexica for 1000+ Languages},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4113--4120},
  abstract  = {In this paper, we introduce UniSent universal sentiment lexica for 1000+ languages. Sentiment lexica are vital for sentiment analysis in absence of document-level annotations, a very common scenario for low-resource languages. To the best of our knowledge, UniSent is the largest sentiment resource to date in terms of the number of covered languages, including many low resource ones. In this work, we use a massively parallel Bible corpus to project sentiment information from English to other languages for sentiment analysis on Twitter data. We introduce a method called DomDrift to mitigate the huge domain mismatch between Bible and Twitter by a confidence weighting scheme that uses domain-specific embeddings to compare the nearest neighbors for a candidate sentiment word in the source (Bible) and target (Twitter) domain. We evaluate the quality of UniSent in a subset of languages for which manually created ground truth was available, Macedonian, Czech, German, Spanish, and French. We show that the quality of UniSent is comparable to manually created sentiment resources when it is used as the sentiment seed for the task of word sentiment prediction on top of embedding representations. In addition, we show that emoticon sentiments could be reliably predicted in the Twitter domain using only UniSent and monolingual embeddings in German, Spanish, French, and Italian. With the publication of this paper, we release the UniSent sentiment lexica at http://language-lab.info/unisent.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.506}
}

@InProceedings{nguyen-bryant:2020:LREC,
  author    = {Nguyen, Li  and  Bryant, Christopher},
  title     = {CanVEC - the Canberra Vietnamese-English Code-switching Natural Speech Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4121--4129},
  abstract  = {This paper introduces the Canberra Vietnamese-English Code-switching corpus (CanVEC), an original corpus of natural mixed speech that we semi-automatically annotated with language information, part of speech (POS) tags and Vietnamese translations.
```

```
The corpus, which was built to inform a sociolinguistic study on
language variation and code-switching, consists of 10 hours of
recorded speech (87k tokens) between 45 Vietnamese-English
bilinguals living in Canberra, Australia. We describe how we
collected and annotated the corpus by pipelining several monolingual
toolkits to considerably speed up the annotation process. We also
describe how we evaluated the automatic annotations to ensure corpus
reliability. We make the corpus available for research purposes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.507}
}

@InProceedings{eryani-EtAl:2020:LREC,
  author    = {Eryani, Fadhl  and  Habash, Nizar  and  Bouamor,
Houda  and  Khalifa, Salam},
  title     = {A Spelling Correction Corpus for Multiple Arabic
Dialects},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4130--4138},
  abstract  = {Arabic dialects are the non-standard varieties of
Arabic commonly spoken -- and increasingly written on social media
-- across the Arab world. Arabic dialects do not have standard
orthographies, a challenge for natural language processing
applications. In this paper, we present the MADAR CODA Corpus, a
collection of 10,000 sentences from five Arabic city dialects
(Beirut, Cairo, Doha, Rabat, and Tunis) represented in the
Conventional Orthography for Dialectal Arabic (CODA) in parallel
with their raw original form. The sentences come from the Multi-
Arabic Dialect Applications and Resources (MADAR) Project and are in
parallel across the cities (2,000 sentences from each city). This
publicly available resource is intended to support research on
spelling correction and text normalization for Arabic dialects. We
present results on a bootstrapping technique we use to speed up the
CODA annotation, as well as on the degree of similarity across the
dialects before and after CODA annotation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.508}
}

@InProceedings{mutuvi-EtAl:2020:LREC,
  author    = {Mutuvi, Stephen  and  Doucet, Antoine  and  Lejeune,
Gael  and  Odeo, Moses},
  title     = {A Dataset for Multi-lingual Epidemiological Event
Extraction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4139--4144},
```

  abstract = {This paper proposes a corpus for the development and evaluation of tools and techniques for identifying emerging infectious disease threats in online news text. The corpus can not only be used for information extraction, but also for other natural language processing (NLP) tasks such as text classification. We make use of articles published on the Program for Monitoring Emerging Diseases (ProMED) platform, which provides current information about outbreaks of infectious diseases globally. Among the key pieces of information present in the articles is the uniform resource locator (URL) to the online news sources where the outbreaks were originally reported. We detail the procedure followed to build the dataset, which includes leveraging the source URLs to retrieve the news reports and subsequently pre-processing the retrieved documents. We also report on experimental results of event extraction on the dataset using the Data Analysis for Information Extraction in any Language(DAnIEL) system. DAnIEL is a multilingual news surveillance system that leverages unique attributes associated with news reporting to extract events: repetition and saliency. The system has wide geographical and language coverage, including low-resource languages. In addition, we compare different classification approaches in terms of their ability to differentiate between epidemic-related and unrelated news articles that constitute the corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.509}
}

@InProceedings{lin-EtAl:2020:LREC1,
  author   = {Lin, Tzu-Hsiang and Rudnicky, Alexander and Bui, Trung and Kim, Doo Soon and Oh, Jean},
  title   = {Adjusting Image Attributes of Localized Regions with Low-level Dialogue},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {405--412},
  abstract = {Natural Language Image Editing (NLIE) aims to use natural language instructions to edit images. Since novices are inexperienced with image editing techniques, their instructions are often ambiguous and contain high-level abstractions which require complex editing steps. Motivated by this inexperience aspect, we aim to smooth the learning curve by teaching the novices to edit images using low-level command terminologies. Towards this end, we develop a task-oriented dialogue system to investigate low-level instructions for NLIE. Our system grounds language on the level of edit operations, and suggests options for users to choose from. Though compelled to express in low-level terms, user evaluation shows that 25\% of users found our system easy-to-use, resonating with our motivation. Analysis shows that users generally adapt to utilizing the proposed low-level language interface. We also identified object segmentation as the key factor to user satisfaction. Our work demonstrates advantages of low-level, direct

language-action mapping approach that can be applied to other problem domains beyond image editing such as audio editing or industrial design.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.51}
}

@InProceedings{krasselt-EtAl:2020:LREC,
  author   = {Krasselt, Julia and Dressen, Philipp and Fluor, Matthias and Mahlow, Cerstin and Rothenhäusler, Klaus and Runte, Maren},
  title   = {Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4145--4151},
  abstract = {The Swiss Web Corpus for Applied Linguistics (Swiss-AL) is a multilingual (German, French, Italian) collection of texts from selected web sources. Unlike most other web corpora it is not intended for NLP purposes, but rather designed to support data-based and data-driven research on societal and political discourses in Switzerland. It currently contains 8 million texts (approx. 1.55 billion tokens), including news and specialist publications, governmental opinions, and parliamentary records, web sites of political parties, companies, and universities, statements from industry associations and NGOs, etc. A flexible processing pipeline using state-of-the-art components allows researchers in applied linguistics to create tailor-made subcorpora for studying discourse in a wide range of domains. So far, Swiss-AL has been used successfully in research on Swiss public discourses on energy and on antibiotic resistance.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.510}
}

@InProceedings{tachbelie-abate-schultz:2020:LREC,
  author   = {Tachbelie, Martha Yifiru and Abate, Solomon Teferra and Schultz, Tanja},
  title   = {Analysis of GlobalPhone and Ethiopian Languages Speech Corpora for Multilingual ASR},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4152--4156},
  abstract = {In this paper, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta). The aim of the analysis is to select speech data from GP for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. To this end,

phonetic overlaps among GP and Ethiopian languages have been analyzed. The result of our analysis shows that there is much phonetic overlap among Ethiopian languages although they are from three different language families. From GP, Turkish, Uyghur and Croatian are found to have much overlap with the Ethiopian languages. On the other hand, Korean has less phonetic overlap with the rest of the languages. Moreover, morphological complexity of the GP and Ethiopian languages, reflected by type to token ration (TTR) and out of vocabulary (OOV) rate, has been analyzed. Both metrics indicated the morphological complexity of the languages. Korean and Amharic have been identified as extremely morphologically complex compared to the other languages. Tigrigna, Russian, Turkish, Polish, etc. are also among the morphologically complex languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.511}
}

@InProceedings{chen-kageura:2020:LREC,
  author   = {Chen, Long-Huei and Kageura, Kyo},
  title   = {Multilingualization of Medical Terminology: Semantic and Structural Embedding Approaches},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4157--4166},
  abstract = {The multilingualization of terminology is an essential step in the translation pipeline, to ensure the correct transfer of domain-specific concepts. Many institutions and language service providers construct and maintain multilingual terminologies, which constitute important assets. However, the curation of such multilingual resources requires significant human effort; though automatic multilingual term extraction methods have been proposed so far, they are of limited success as term translation cannot be satisfied by simply conveying meaning, but requires the terminologists and domain experts' knowledge to fit the term within the existing terminology. Here we propose a method to encode the structural property of a term by aligning their embeddings using graph convolutional networks trained from separate languages. We observe that the structural information can augment the semantic methods also explored in this work, and recognize the unique nature of terminologies allows our method to fully take advantage and produce superior results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.512}
}

@InProceedings{abate-EtAl:2020:LREC,
  author   = {Abate, Solomon Teferra and Tachbelie, Martha Yifiru and Melese, Michael and Abera, Hafte and Abebe, Tewodros and Mulugeta, Wondwossen and Assabie, Yaregal and Meshesha, Million and Afnafu, Solomon and Seyoum, Binyam Ephrem},
  title   = {Large Vocabulary Read Speech Corpora for Four Ethiopian Languages: Amharic, Tigrigna, Oromo and Wolaytta},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4167--4171},
  abstract  = {Automatic Speech Recognition (ASR) is one of the most
important technologies to support spoken communication in modern
life. However, its development benefits from large speech corpus.
The development of such a corpus is expensive and most of the human
languages, including the Ethiopian languages, do not have such
resources. To address this problem, we have developed four large
(about 22 hours) speech corpora for four Ethiopian languages:
Amharic, Tigrigna, Oromo and Wolaytta. To assess usability of the
corpora for (the purpose of) speech processing, we have developed
ASR systems for each language. In this paper, we present the corpora
and the baseline ASR systems we have developed. We have achieved
word error rates (WERs) of 37.65\%, 31.03\%, 38.02\%, 33.89\% for
Amharic, Tigrigna, Oromo and Wolaytta, respectively. This results
show that the corpora are suitable for further investigation towards
the development of ASR systems. Thus, the research community can use
the corpora to further improve speech processing systems. From our
results, it is clear that the collection of text corpora to train
strong language models for all of the languages is still required,
especially for Oromo and Wolaytta.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.513}
}

@InProceedings{firdaus-ekbal-bhattacharyya:2020:LREC,
  author    = {Firdaus, Mauajama  and  Ekbal, Asif  and
Bhattacharyya, Pushpak},
  title     = {Incorporating Politeness across Languages in Customer
Care Responses: Towards building a Multi-lingual Empathetic Dialogue
Agent},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4172--4182},
  abstract  = {Customer satisfaction is an essential aspect of
customer care systems. It is imperative for such systems to be
polite while handling customer requests/demands. In this paper, we
present a large multi-lingual conversational dataset for English and
Hindi. We choose data from Twitter having both generic and courteous
responses between customer care agents and aggrieved users. We also
propose strong baselines that can induce courteous behaviour in
generic customer care response in a multi-lingual scenario. We build
a deep learning framework that can simultaneously handle different
languages and incorporate polite behaviour in the customer care
agent's responses. Our system is competent in generating responses
in different languages (here, English and Hindi) depending on the
```

customer’s preference and also is able to converse with humans in an empathetic manner to ensure customer satisfaction and retention. Experimental results show that our proposed models can converse in both the languages and the information shared between the languages helps in improving the performance of the overall system. Qualitative and quantitative analysis shows that the proposed method can converse in an empathetic manner by incorporating courteousness in the responses and hence increasing customer satisfaction.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.514}
}

@InProceedings{sas-beloucif-sgaard:2020:LREC,
  author   = {Sas, Cezar and Beloucif, Meriem and Søgaard, Anders},
  title   = {WikiBank: Using Wikidata to Improve Multilingual Frame-Semantic Parsing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4183--4189},
  abstract = {Frame-semantic annotations exist for a tiny fraction of the world's languages, Wikidata, however, links knowledge base triples to texts in many languages, providing a common, distant supervision signal for semantic parsers. We present WikiBank, a multilingual resource of partial semantic structures that can be used to extend pre-existing resources rather than creating new man-made resources from scratch. We also integrate this form of supervision into an off-the-shelf frame-semantic parser and allow cross-lingual transfer. Using Google's Sling architecture, we show significant improvements on the English and Spanish CoNLL 2009 datasets, whether training on the full available datasets or small subsamples thereof.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.515}
}

@InProceedings{ahmed-EtAl:2020:LREC1,
  author   = {Ahmed, Mahtab and Dixit, Chahna and Mercer, Robert E. and Khan, Atif and Samee, Muhammad Rifayat and Urra, Felipe},
  title   = {Multilingual Corpus Creation for Multilingual Semantic Similarity Task},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4190--4196},
  abstract = {In natural language processing, the performance of a semantic similarity task relies heavily on the availability of a large corpus. Various monolingual corpora are available (mainly

English); but multilingual resources are very limited. In this work, we describe a semi-automated framework to create a multilingual corpus which can be used for the multilingual semantic similarity task. The similar sentence pairs are obtained by crawling bilingual websites, whereas the dissimilar sentence pairs are selected by applying topic modeling and an Open-AI GPT model on the similar sentence pairs. We focus on websites in the government, insurance, and banking domains to collect English-French and English-Spanish sentence pairs; however, this corpus creation approach can be applied to any other industry vertical provided that a bilingual website exists. We also show experimental results for multilingual semantic similarity to verify the quality of the corpus and demonstrate its usage.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.516}
}

@InProceedings{wang-EtAl:2020:LREC1,
  author   = {Wang, Changhan and Pino, Juan and Wu, Anne and Gu, Jiatao},
  title   = {CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4197--4203},
  abstract = {Spoken language translation has recently witnessed a resurgence in popularity, thanks to the development of end-to-end models and the creation of new corpora, such as Augmented LibriSpeech and MuST-C. Existing datasets involve language pairs with English as a source language, involve very specific domains or are low resource. We introduce CoVoST, a multilingual speech-to-text translation corpus from 11 languages into English, diversified with over 11,000 speakers and over 60 accents. We describe the dataset creation methodology and provide empirical evidence of the quality of the data. We also provide initial benchmarks, including, to our knowledge, the first end-to-end many-to-one multilingual models for spoken language translation. CoVoST is released under CC0 license and free to use. We also provide additional evaluation data derived from Tatoeba under CC licenses.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.517}
}

@InProceedings{nakayama-tamura-ninomiya:2020:LREC,
  author   = {Nakayama, Hideki and Tamura, Akihiro and Ninomiya, Takashi},
  title   = {A Visually-Grounded Parallel Corpus with Phrase-to-Region Linking},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4204--4210},
  abstract  = {Visually-grounded natural language processing has
become an important research direction in the past few years.
However, majorities of the available cross-modal resources (e.g.,
image-caption datasets) are built in English and cannot be directly
utilized in multilingual or non-English scenarios. In this study, we
present a novel multilingual multimodal corpus by extending the
Flickr30k Entities image-caption dataset with Japanese translations,
which we name Flickr30k Entities JP (F30kEnt-JP). To the best of our
knowledge, this is the first multilingual image-caption dataset
where the captions in the two languages are parallel and have the
shared annotations of many-to-many phrase-to-region linking. We
believe that phrase-to-region as well as phrase-to-phrase
supervision can play a vital role in fine-grained grounding of
language and vision, and will promote many tasks such as
multilingual image captioning and multimodal machine translation. To
verify our dataset, we performed phrase localization experiments in
both languages and investigated the effectiveness of our Japanese
annotations as well as multilingual learning realized by our
dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.518}
}

@InProceedings{wu-nicolai-yarowsky:2020:LREC,
  author    = {Wu, Winston  and  Nicolai, Garrett  and  Yarowsky,
David},
  title     = {Multilingual Dictionary Based Construction of Core
Vocabulary},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4211--4217},
  abstract  = {We propose a new functional definition and
construction method for core vocabulary sets for multiple
applications based on the relative coverage of a target concept in
thousands of bilingual dictionaries. Our newly developed core
concept vocabulary list derived from these dictionary consensus
methods achieves high overlap with existing widely utilized core
vocabulary lists targeted at applications such as first and second
language learning or field linguistics. Our in-depth analysis
illustrates multiple desirable properties of our newly proposed core
vocabulary set, including their non-compositionality. We employ a
cognate prediction method to recover missing coverage of this core
vocabulary in massively multilingual dictionary construction, and we
argue that this core vocabulary should be prioritized for
elicitation when creating new dictionaries for low-resource
languages for multiple downstream tasks including machine
translation and language learning.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.519}
```

```
}

@InProceedings{yim-EtAl:2020:LREC,
  author    = {Yim, Wen-wai  and  Yetisgen, Meliha  and  Huang,
Jenny  and  Grossman, Micah},
  title     = {Alignment Annotation for Clinic Visit Dialogue to
Clinical Note Sentence Language Generation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {413--421},
  abstract  = {For every patient's visit to a clinician, a clinical
note is generated documenting their medical conversation, including
complaints discussed, treatments, and medical plans. Despite
advances in natural language processing, automating clinical note
generation from a clinic visit conversation is a largely unexplored
area of research. Due to the idiosyncrasies of the task, traditional
methods of corpus creation are not effective enough approaches for
this problem. In this paper, we present an annotation methodology
that is content- and technique- agnostic while associating note
sentences to sets of dialogue sentences. The sets can further be
grouped with higher order tags to mark sets with related
information. This direct linkage from input to output decouples the
annotation from specific language understanding or generation
strategies. Here we provide data statistics and qualitative analysis
describing the unique annotation challenges. Given enough annotated
data, such a resource would support multiple modeling methods
including information extraction with template language generation,
information retrieval type language generation, or sequence to
sequence modeling.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.52}
}

@InProceedings{ardila-EtAl:2020:LREC,
  author    = {Ardila, Rosana  and  Branson, Megan  and  Davis,
Kelly  and  Kohler, Michael  and  Meyer, Josh  and  Henretty,
Michael  and  Morais, Reuben  and  Saunders, Lindsay  and  Tyers,
Francis  and  Weber, Gregor},
  title     = {Common Voice: A Massively-Multilingual Speech
Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4218--4222},
  abstract  = {The Common Voice corpus is a massively-multilingual
collection of transcribed speech intended for speech technology
research and development. Common Voice is designed for Automatic
Speech Recognition purposes but can be useful in other domains (e.g.
```

language identification). To achieve scale and sustainability, the Common Voice project employs crowdsourcing for both data collection and data validation. The most recent release includes 29 languages, and as of November 2019 there are a total of 38 languages collecting data. Over 50,000 individuals have participated so far, resulting in 2,500 hours of collected audio. To our knowledge this is the largest audio corpus in the public domain for speech recognition, both in terms of number of hours and number of languages. As an example use case for Common Voice, we present speech recognition experiments using Mozilla's DeepSpeech Speech-to-Text toolkit. By applying transfer learning from a source English model, we find an average Character Error Rate improvement of 5.99 ± 5.48 for twelve target languages (German, French, Italian, Turkish, Catalan, Slovenian, Welsh, Irish, Breton, Tatar, Chuvash, and Kabyle). For most of these languages, these are the first ever published results on end-to-end Automatic Speech Recognition.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.520}
}

@InProceedings{lee-EtAl:2020:LREC1,
  author   = {Lee, Jackson L. and Ashby, Lucas F.E. and Garza, M. Elizabeth and Lee-Sikka, Yeonju and Miller, Sean and Wong, Alan and McCarthy, Arya D. and Gorman, Kyle},
  title   = {Massively Multilingual Pronunciation Modeling with WikiPron},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4223--4228},
  abstract = {We introduce WikiPron, an open-source command-line tool for extracting pronunciation data from Wiktionary, a collaborative multilingual online dictionary. We first describe the design and use of WikiPron. We then discuss the challenges faced scaling this tool to create an automatically-generated database of 1.7 million pronunciations from 165 languages. Finally, we validate the pronunciation database by using it to train and evaluating a collection of generic grapheme-to-phoneme models. The software, pronunciation data, and models are all made available under permissive open-source licenses.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.521}
}

@InProceedings{ylijyr-EtAl:2020:LREC,
  author   = {Yli-Jyrä, Anssi and Purhonen, Josi and Liljeqvist, Matti and Antturi, Arto and Nieminen, Pekka and Räntilä, Kari M. and Luoto, Valtter},
  title   = {HELFI: a Hebrew-Greek-Finnish Parallel Bible Corpus with Cross-Lingual Morpheme Alignment},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {4229--4236},
  abstract  = {Twenty-five years ago, morphologically aligned
Hebrew-Finnish and Greek-Finnish bitexts (texts accompanied by a
translation) were constructed manually in order to create an
analytical concordance (Luoto et al., eds. 1997) for a Finnish Bible
translation. The creators of the bitexts recently secured the
publisher's permission to release its fine-grained alignment, but
the alignment was still dependent on proprietary, third-party
resources such as a copyrighted text edition and proprietary
morphological analyses of the source texts. In this paper, we
describe a nontrivial editorial process starting from the creation
of the original one-purpose database and ending with its
reconstruction using only freely available text editions and
annotations. This process produced an openly available dataset that
contains (i) the source texts and their translations, (ii) the
morphological analyses, (iii) the cross-lingual morpheme
alignments.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.522}
}

@InProceedings{hamed-vu-abdennadher:2020:LREC,
  author    = {Hamed, Injy  and  Vu, Ngoc Thang  and  Abdennadher,
Slim},
  title     = {ArzEn: A Speech Corpus for Code-switched Egyptian
Arabic-English},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4237--4246},
  abstract  = {In this paper, we present our ArzEn corpus, an
Egyptian Arabic-English code-switching (CS) spontaneous speech
corpus. The corpus is collected through informal interviews with 38
Egyptian bilingual university students and employees held in a
soundproof room. A total of 12 hours are recorded, transcribed,
validated and sentence segmented. The corpus is mainly designed to
be used in Automatic Speech Recognition (ASR) systems, however, it
also provides a useful resource for analyzing the CS phenomenon from
linguistic, sociological, and psychological perspectives. In this
paper, we first discuss the CS phenomenon in Egypt and the factors
that gave rise to the current language. We then provide a detailed
description on how the corpus was collected, giving an overview on
the participants involved. We also present statistics on the CS
involved in the corpus, as well as a summary to the effort exerted
in the corpus development, in terms of number of hours required for
transcription, validation, segmentation and speaker annotation.
Finally, we discuss some factors contributing to the complexity of
the corpus, as well as Arabic-English CS behaviour that could pose
potential challenges to ASR systems.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.523}
}

@InProceedings{khakhmovich-EtAl:2020:LREC,
  author    = {Khakhmovich, Aleksandr  and  Pavlova, Svetlana  and
Kirillova, Kira  and  Arefyev, Nikolay  and  Savilova, Ekaterina},
  title     = {Cross-lingual Named Entity List Search via
Transliteration},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4247--4255},
  abstract  = {Out-of-vocabulary words are still a challenge in
cross-lingual Natural Language Processing tasks, for which
transliteration from source to target language or script is one of
the solutions. In this study, we collect a personal name dataset in
445 Wikidata languages (37 scripts), train Transformer-based
multilingual transliteration models on 6 high- and 4 less-resourced
languages, compare them with bilingual models from (Merhav and Ash,
2018) and determine that multilingual models perform better for
less-resourced languages. We discover that intrinsic evaluation, i.e
comparison to a single gold standard, might not be appropriate in
the task of transliteration due to its high variability. For this
reason, we propose using extrinsic evaluation of transliteration via
the cross-lingual named entity list search task (e.g. personal name
search in contacts list). Our code and datasets are publicly
available online.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.524}
}

@InProceedings{bost-labatut-linares:2020:LREC,
  author    = {Bost, Xavier  and  Labatut, Vincent  and  Linares,
Georges},
  title     = {Serial Speakers: a Dataset of TV Series},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4256--4264},
  abstract  = {For over a decade, TV series have been drawing
increasing interest, both from the audience and from various
academic fields. But while most viewers are hooked on the continuous
plots of TV serials, the few annotated datasets available to
researchers focus on standalone episodes of classical TV series. We
aim at filling this gap by providing the multimedia/speech
processing communities with ``Serial Speakers'', an annotated
dataset of 155 episodes from three popular American TV serials:
``Breaking Bad'', ``Game of Thrones'' and ``House of Cards''.
``Serial Speakers'' is suitable both for investigating multimedia
```

retrieval in realistic use case scenarios, and for addressing lower level speech related tasks in especially challenging conditions. We publicly release annotations for every speech turn (boundaries, speaker) and scene boundary, along with annotations for shot boundaries, recurring shots, and interacting speakers in a subset of episodes. Because of copyright restrictions, the textual content of the speech turns is encrypted in the public version of the dataset, but we provide the users with a simple online tool to recover the plain text from their own subtitle files.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.525}
}

@InProceedings{muraoka-kohita-ishii:2020:LREC,
  author   = {Muraoka, Masayasu and Kohita, Ryosuke and Ishii, Etsuko},
  title   = {Image Position Prediction in Multimodal Documents},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4265--4274},
  abstract = {Conventional multimodal tasks, such as caption generation and visual question answering, have allowed machines to understand an image by describing or being asked about it in natural language, often via a sentence. Datasets for these tasks contain a large number of pairs of an image and the corresponding sentence as an instance. However, a real multimodal document such as a news article or Wikipedia page consists of multiple sentences with multiple images. Such documents require an advanced skill of jointly considering the multiple texts and multiple images, beyond a single sentence and image, for the interpretation. Therefore, aiming at building a system that can understand multimodal documents, we propose a task called image position prediction (IPP). In this task, a system learns plausible positions of images in a given document. To study this task, we automatically constructed a dataset of 66K multimodal documents with 320K images from Wikipedia articles. We conducted a preliminary experiment to evaluate the performance of a current multimodal system on our task. The experimental results show that the system outperformed simple baselines while the performance is still far from human performance, which thus poses new challenges in multimodal research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.526}
}

@InProceedings{nishimura-EtAl:2020:LREC,
  author   = {Nishimura, Taichi and Tomori, Suzushi and Hashimoto, Hayato and Hashimoto, Atsushi and Yamakata, Yoko and Harashima, Jun and Ushiku, Yoshitaka and Mori, Shinsuke},
  title   = {Visual Grounding Annotation of Recipe Flow Graph},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {4275--4284},
  abstract  = {In this paper, we provide a dataset that gives visual
grounding annotations to recipe flow graphs. A recipe flow graph is
a representation of the cooking workflow, which is designed with the
aim of understanding the workflow from natural language processing.
Such a workflow will increase its value when grounded to real-world
activities, and visual grounding is a way to do so. Visual grounding
is provided as bounding boxes to image sequences of recipes, and
each bounding box is linked to an element of the workflow. Because
the workflows are also linked to the text, this annotation gives
visual grounding with workflow's contextual information between
procedural text and visual observation in an indirect manner. We
subsidiarily annotated two types of event attributes with each
bounding box: ``doing-the-action,'' or ``done-the-action''. As a
result of the annotation, we got 2,300 bounding boxes in 272 flow
graph recipes. Various experiments showed that the proposed dataset
enables us to estimate contextual information described in recipe
flow graphs from an image sequence.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.527}
}

@InProceedings{adjali-EtAl:2020:LREC,
  author    = {Adjali, Omar  and  Besançon, Romaric  and  Ferret,
Olivier  and  Le Borgne, Hervé  and  Grau, Brigitte},
  title     = {Building a Multimodal Entity Linking Dataset From
Tweets},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4285--4292},
  abstract  = {The task of Entity linking, which aims at associating
an entity mention with a unique entity in a knowledge base (KB), is
useful for advanced Information Extraction tasks such as relation
extraction or event detection. Most of the studies that address this
problem rely only on textual documents while an increasing number of
sources are multimedia, in particular in the context of social media
where messages are often illustrated with images. In this article,
we address the Multimodal Entity Linking (MEL) task, and more
particularly the problem of its evaluation. To this end, we propose
a novel method to quasi-automatically build annotated datasets to
evaluate methods on the MEL task. The method collects text and
images to jointly build a corpus of tweets with ambiguous mentions
along with a Twitter KB defining the entities. We release a new
annotated dataset of Twitter posts associated with images. We study
the key characteristics of the proposed dataset and evaluate the
performance of several MEL approaches on it.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.528}
}
```

```
@InProceedings{mdhaffar-EtAl:2020:LREC,
  author    = {mdhaffar, salima and Estève, Yannick and Laurent, Antoine and Hernandez, Nicolas and Dufour, Richard and Charlet, Delphine and Damnati, Geraldine and Quiniou, Solen and Camelin, Nathalie},
  title     = {A Multimodal Educational Corpus of Oral Courses: Annotation, Analysis and Case Study},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4293--4301},
  abstract  = {This corpus is part of the PASTEL (Performing Automated Speech Transcription for Enhancing Learning) project aiming to explore the potential of synchronous speech transcription and application in specific teaching situations. It includes 10 hours of different lectures, manually transcribed and segmented. The main interest of this corpus lies in its multimodal aspect: in addition to speech, the courses were filmed and the written presentation supports (slides) are made available. The dataset may then serve researches in multiple fields, from speech and language to image and video processing. The dataset will be freely available to the research community. In this paper, we first describe in details the annotation protocol, including a detailed analysis of the manually labeled data. Then, we propose some possible use cases of the corpus with baseline results. The use cases concern scientific fields from both speech and text processing, with language model adaptation, thematic segmentation and transcription to slide alignment.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.529}
}

@InProceedings{eric-EtAl:2020:LREC,
  author    = {Eric, Mihail and Goel, Rahul and Paul, Shachi and Sethi, Abhishek and Agarwal, Sanchit and Gao, Shuyang and Kumar, Adarsh and Goyal, Anuj and Ku, Peter and Hakkani-Tur, Dilek},
  title     = {MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {422--428},
  abstract  = {MultiWOZ 2.0 (Budzianowski et al., 2018) is a recently released multi-domain dialogue dataset spanning 7 distinct domains and containing over 10,000 dialogues. Though immensely useful and one of the largest resources of its kind to-date, MultiWOZ 2.0 has a few shortcomings. Firstly, there are substantial
```

noise in the dialogue state annotations and dialogue utterances which negatively impact the performance of state-tracking models. Secondly, follow-up work (Lee et al., 2019) has augmented the original dataset with user dialogue acts. This leads to multiple co-existent versions of the same dataset with minor modifications. In this work we tackle the aforementioned issues by introducing MultiWOZ 2.1. To fix the noisy state annotations, we use crowdsourced workers to re-annotate state and utterances based on the original utterances in the dataset. This correction process results in changes to over 32\% of state annotations across 40\% of the dialogue turns. In addition, we fix 146 dialogue utterances by canonicalizing slot values in the utterances to the values in the dataset ontology. To address the second problem, we combined the contributions of the follow-up works into MultiWOZ 2.1. Hence, our dataset also includes user dialogue acts as well as multiple slot descriptions per dialogue state slot. We then benchmark a number of state-of-the-art dialogue state tracking models on the MultiWOZ 2.1 dataset and show the joint state tracking performance on the corrected state annotations. We are publicly releasing MultiWOZ 2.1 to the community, hoping that this dataset resource will allow for more effective models across various dialogue subproblems to be built in the future.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.53}
}

@InProceedings{kameko-mori:2020:LREC,
  author   = {Kameko, Hirotaka and Mori, Shinsuke},
  title   = {Annotating Event Appearance for Japanese Chess Commentary Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4302--4308},
  abstract = {In recent years, there has been a surge of interest in natural language processing related to the real world, such as symbol grounding, language generation, and non-linguistic data search by natural language queries. Researchers usually collect pairs of text and non-text data for research. However, the text and non-text data are not always a “true” pair. We focused on the shogi (Japanese chess) commentaries, which are accompanied by game states as a well-defined “real world”. For analyzing and processing texts accurately, considering only the given states is insufficient, and we must consider the relationship between texts and the real world. In this paper, we propose “Event Appearance” labels that show the relationship between events mentioned in texts and those happening in the real world. Our event appearance label set consists of temporal relation, appearance probability, and evidence of the event. Statistics of the annotated corpus and the experimental result show that there exists temporal relation which skillful annotators realize in common. However, it is hard to predict the relationship only by considering the given states.},

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.530}
}

@InProceedings{alcntara-moreira-feijo:2020:LREC,
  author    = {Alcântara, Cleber  and  Moreira, Viviane  and  Feijo,
Diego},
  title     = {Offensive Video Detection: Dataset and Baseline
Results},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4309--4319},
  abstract  = {Web-users produce and publish high volumes of data of
various types, such as text, images, and videos. The platforms try
to restrain their users from publishing offensive content to keep a
friendly and respectful environment and rely on moderators to filter
the posts. However, this method is insufficient due to the high
volume of publications. The identification of offensive material can
be performed automatically using machine learning, which needs
annotated datasets. Among the published datasets in this matter, the
Portuguese language is underrepresented, and videos are little
explored. We investigated the problem of offensive video detection
by assembling and publishing a dataset of videos in Portuguese
containing mostly textual features. We ran experiments using popular
machine learning classifiers used in this domain and reported our
findings, alongside multiple evaluation metrics. We found that using
word embedding with Deep Learning classifiers achieved the best
results on average. CNN architectures, Naive Bayes, and Random
Forest ranked top among different experiments. Transfer Learning
models outperformed Classic algorithms when processing video
transcriptions, but scored lower using other feature sets. These
findings can be used as a baseline for future works on this
subject.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.531}
}

@InProceedings{trotta-EtAl:2020:LREC,
  author    = {Trotta, Daniela  and  Palmero Aprosio, Alessio  and
Tonelli, Sara  and  Elia, Annibale},
  title     = {Adding Gesture, Posture and Facial Displays to the
PoliModal Corpus of Political Interviews},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4320--4326},
  abstract  = {This paper introduces a multimodal corpus in the
political domain, which on top of transcribed face-to-face
interviews presents the annotation of facial displays, hand gestures
```

and body posture. While the fully annotated corpus consists of 3 interviews for a total of 90 minutes, it is extracted from a larger available corpus of 56 face-to-face interviews (14 hours) that has been manually annotated with information about metadata (i.e. tools used for the transcription, link to the interview etc.), pauses (used to mark a pause either between or within utterances), vocal expressions (marking non-lexical expressions such as burp and semi-lexical expressions such as primary interjections), deletions (false starts, repetitions and truncated words) and overlaps. In this work, we describe the additional level of annotation relating to nonverbal elements used by three Italian politicians belonging to three different political parties and who at the time of the talk-show were all candidates for the presidency of the Council of Minister. We also present the results of some analyses aimed at identifying existing relations between the proxemics phenomena and the linguistic structures in which they occur in order to capture recurring patterns and differences in the communication strategy.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.532}
}

@InProceedings{hodac-fleury-ponton:2020:LREC,
  author   = {Ho-Dac, Lydia-Mai and Fleury, Serge and Ponton, Claude},
  title   = {E:Calm Resource: a Resource for Studying Texts Produced by French Pupils and Students},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4327--4332},
  abstract = {The E:Calm resource is constructed from French student texts produced in a variety of usual contexts of teaching. The distinction of the E:Calm resource is to provide an ecological data set that gives a broad overview of texts written at elementary school, high school and university. This paper describes the whole data processing: encoding of the main graphical aspects of the handwritten primary sources according to the TEI-P5 norm; spelling standardizing; POS tagging and syntactic parsing evaluation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.533}
}

@InProceedings{jansen-EtAl:2020:LREC,
  author   = {Jansen, Michel-Pierre and Truong, Khiet P. and Heylen, Dirk K.J. and Nazareth, Deniece S.},
  title   = {Introducing MULAI: A Multimodal Database of Laughter during Dyadic Interactions},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {4333--4342},
  abstract  = {Although laughter has gained considerable interest
from a diversity of research areas, there still is a need for
laughter specific databases. We present the Multimodal Laughter
during Interaction (MULAI) database to study the expressive patterns
of conversational and humour related laughter. The MULAI database
contains 2 hours and 14 minutes of recorded and annotated dyadic
human-human interactions and includes 601 laughs, 168 speech-laughs
and 538 on- or offset respirations. This database is unique in
several ways; 1) it focuses on different types of social laughter
including conversational- and humour related laughter, 2) it
contains annotations from participants, who understand the social
context, on how humourous they perceived themselves and their
interlocutor during each task, and 3) it contains data rarely
captured by other laughter databases including participant
personality profiles and physiological responses. We use the MULAI
database to explore the link between acoustic laughter properties
and annotated humour ratings over two settings. The results reveal
that the duration, pitch and intensity of laughs from participants
do not correlate with their own perception of how humourous they
are, however the acoustics of laughter do correlate with how
humourous they are being perceived by their conversational
partner.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.534}
}

@InProceedings{oostdijk-EtAl:2020:LREC,
  author    = {Oostdijk, Nelleke  and  van Halteren, Hans  and
Bașar, Erkan  and  Larson, Martha},
  title     = {The Connection between the Text and Images of News
Articles: New Insights for Multimedia Analysis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4343--4351},
  abstract  = {We report on a case study of text and images that
reveals the inadequacy of simplistic assumptions about their
connection and interplay. The context of our work is a larger effort
to create automatic systems that can extract event information from
online news articles about flooding disasters. We carry out a manual
analysis of 1000 articles containing a keyword related to flooding.
The analysis reveals that the articles in our data set cluster into
seven categories related to different topical aspects of flooding,
and that the images accompanying the articles cluster into five
categories related to the content they depict. The results
demonstrate that flood-related news articles do not consistently
report on a single, currently unfolding flooding event and we should
also not assume that a flood-related image will directly relate to a
flooding-event described in the corresponding article. In
particular, spatiotemporal distance is important. We validate the
manual analysis with an automatic classifier demonstrating the
```

```
technical feasibility of multimedia analysis approaches that admit
more realistic relationships between text and images. In sum, our
case study confirms that closer attention to the connection between
text and images has the potential to improve the collection of
multimodal information from news articles.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.535}
}

@InProceedings{castro-EtAl:2020:LREC,
  author    = {Castro, Santiago  and  Azab, Mahmoud  and  Stroud,
Jonathan  and  Noujaim, Cristina  and  Wang, Ruoyao  and  Deng, Jia
and  Mihalcea, Rada},
  title     = {LifeQA: A Real-life Dataset for Video Question
Answering},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4352--4358},
  abstract  = {We introduce LifeQA, a benchmark dataset for video
question answering that focuses on day-to-day real-life situations.
Current video question answering datasets consist of movies and TV
shows. However, it is well-known that these visual domains are not
representative of our day-to-day lives. Movies and TV shows, for
example, benefit from professional camera movements, clean editing,
crisp audio recordings, and scripted dialog between professional
actors. While these domains provide a large amount of data for
training models, their properties make them unsuitable for testing
real-life question answering systems. Our dataset, by contrast,
consists of video clips that represent only real-life scenarios. We
collect 275 such video clips and over 2.3k multiple-choice
questions. In this paper, we analyze the challenging but realistic
aspects of LifeQA, and we apply several state-of-the-art video
question answering models to provide benchmarks for future research.
The full dataset is publicly available at https://
lit.eecs.umich.edu/lifeqa/.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.536}
}

@InProceedings{bettinger-EtAl:2020:LREC,
  author    = {Bettinger, Julia  and  Hätty, Anna  and  Dorna,
Michael  and  Schulte im Walde, Sabine},
  title     = {A Domain-Specific Dataset of Difficulty Ratings for
German Noun Compounds in the Domains DIY, Cooking and Automotive},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4359--4367},
  abstract  = {We present a dataset with difficulty ratings for
```

```
1,030 German closed noun compounds extracted from domain-specific
texts for do-it-ourself (DIY), cooking and automotive. The dataset
includes two-part compounds for cooking and DIY, and two- to four-
part compounds for automotive. The compounds were identified in text
using the Simple Compound Splitter (Weller-Di Marco, 2017); a subset
was filtered and balanced for frequency and productivity criteria as
basis for manual annotation and fine-grained interpretation. This
study presents the creation, the final dataset with ratings from 20
annotators and statistics over the dataset, to provide insight into
the perception of domain-specific term difficulty. It is
particularly striking that annotators agree on a coarse, binary
distinction between easy vs. difficult domain-specific compounds but
that a more fine grained distinction of difficulty is not
meaningful. We finally discuss the challenges of an annotation for
difficulty, which includes both the task description as well as the
selection of the data basis.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.537}
}

@InProceedings{strakatova-EtAl:2020:LREC,
  author    = {Strakatova, Yana  and  Falk, Neele  and  Fuhrmann,
Isabel  and  Hinrichs, Erhard  and  Rossmann, Daniela},
  title     = {All That Glitters is Not Gold: A Gold Standard of
Adjective-Noun Collocations for German},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4368--4378},
  abstract  = {In this paper we present the GerCo dataset of
adjective-noun collocations for German, such as alter Freund `old
friend' and tiefe Liebe `deep love'. The annotation has been
performed by experts based on the annotation scheme introduced in
this paper. The resulting dataset contains 4,732 positive and
negative instances of collocations and covers all the 16 semantic
classes of adjectives as defined in the German wordnet GermaNet. The
dataset can serve as a reliable empirical basis for comparing
different theoretical frameworks concerned with collocations or as
material for data-driven approaches to the studies of collocations
including different machine learning experiments. This paper
addresses the latter issue by using the GerCo dataset for evaluating
different models on the task of automatic collocation
identification. We compare lexical association measures with static
and contextualized word embeddings. The experiments show that word
embeddings outperform methods based on statistical association
measures by a wide margin.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.538}
}

@InProceedings{alipoor-schulteimwalde:2020:LREC,
  author    = {Alipoor, Pegah  and  Schulte im Walde, Sabine},
  title     = {Variants of Vector Space Reductions for Predicting
```

the Compositionality of English Noun Compounds},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4379--4387},
  abstract = {Predicting the degree of compositionality of noun compounds such as "snowball" and "butterfly" is a crucial ingredient for lexicography and Natural Language Processing applications, to know whether the compound should be treated as a whole, or through its constituents, and what it means. Computational approaches for an automatic prediction typically represent and compare compounds and their constituents within a vector space and use distributional similarity as a proxy to predict the semantic relatedness between the compounds and their constituents as the compound’s degree of compositionality. This paper provides a systematic evaluation of vector-space reduction variants across kinds, exploring reductions based on part-of-speech next to and also in combination with Principal Components Analysis using Singular Value and word2vec embeddings. We show that word2vec and nouns only dimensionality reductions are the most successful and stable vector space variants for our task.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.539}
}

@InProceedings{kraus-EtAl:2020:LREC,
  author   = {Kraus, Matthias and Fischbach, Fabian and Jansen, Pascal and Minker, Wolfgang},
  title   = {A Comparison of Explicit and Implicit Proactive Dialogue Strategies for Conversational Recommendation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {429--435},
  abstract = {Recommendation systems aim at facilitating information retrieval for users by taking into account their preferences. Based on previous user behaviour, such a system suggests items or provides information that a user might like or find useful. Nonetheless, how to provide suggestions is still an open question. Depending on the way a recommendation is communicated influences the user’s perception of the system. This paper presents an empirical study on the effects of proactive dialogue strategies on user acceptance. Therefore, an explicit strategy based on user preferences provided directly by the user, and an implicit proactive strategy, using autonomously gathered information, are compared. The results show that proactive dialogue systems significantly affect the perception of human-computer interaction. Although no significant differences are found between implicit and explicit strategies, proactivity significantly influences the user experience

```
compared to reactive system behaviour. The study contributes new
insights to the human-agent interaction and the voice user interface
design. Furthermore, we discover interesting tendencies that
motivate futurework.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.54}
}

@InProceedings{nigam-htty-schulteimwalde:2020:LREC,
  author    = {Nigam, Anurag  and  Hätty, Anna  and  Schulte im
Walde, Sabine},
  title     = {Varying Vector Representations and Integrating
Meaning Shifts into a PageRank Model for Automatic Term Extraction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4388--4394},
  abstract  = {We perform a comparative study for automatic term
extraction from domain-specific language using a PageRank model with
different edge-weighting methods. We vary vector space
representations within the PageRank graph algorithm, and we go
beyond standard co-occurrence and investigate the influence of
measures of association strength and first- vs. second-order co-
occurrence. In addition, we incorporate meaning shifts from general
to domain-specific language as personalized vectors, in order to
distinguish between termhood strengths of ambiguous words across
word senses. Our study is performed for two domain-specific English
corpora: ACL and do-it-yourself (DIY); and a domain-specific German
corpus: cooking. The models are assessed by applying average
precision and the roc score as evaluation metrices.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.540}
}

@InProceedings{fort-EtAl:2020:LREC,
  author    = {Fort, Karën  and  Guillaume, Bruno  and  Pilatte,
Yann-Alan  and  Constant, Mathieu  and  Lefèbvre, Nicolas},
  title     = {Rigor Mortis: Annotating MWEs with a Gamified
Platform},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4395--4401},
  abstract  = {We present here Rigor Mortis, a gamified
crowdsourcing platform designed to evaluate the intuition of the
speakers, then train them to annotate multi-word expressions (MWEs)
in French corpora. We previously showed that the speakers' intuition
is reasonably good (65\% in recall on non-fixed MWE). We detail here
the annotation results, after a training phase using some of the
tests developed in the PARSEME-FR project.},
```

```
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.541}
}

@InProceedings{kurfal-EtAl:2020:LREC,
  author    = {Kurfalı, Murathan  and  Östling, Robert  and  Sjons,
Johan  and  Wirén, Mats},
  title     = {A Multi-word Expression Dataset for Swedish},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4402--4409},
  abstract  = {We present a new set of 96 Swedish multi-word
expressions annotated with degree of (non-)compositionality. In
contrast to most previous compositionality datasets we also consider
syntactically complex constructions and publish a formal
specification of each expression. This allows evaluation of
computational models beyond word bigrams, which have so far been the
norm. Finally, we use the annotations to evaluate a system for
automatic compositionality estimation based on distributional
semantics. Our analysis of the disagreements between human
annotators and the distributional model reveal interesting questions
related to the perception of compositionality, and should be
informative to future work in the area.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.542}
}

@InProceedings{krotova-aksenov-artemova:2020:LREC,
  author    = {Krotova, Irina  and  Aksenov, Sergey  and  Artemova,
Ekaterina},
  title     = {A Joint Approach to Compound Splitting and Idiomatic
Compound Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4410--4417},
  abstract  = {Applications such as machine translation, speech
recognition, and information retrieval require efficient handling of
noun compounds as they are one of the possible sources for out of
vocabulary words. In-depth processing of noun compounds requires not
only splitting them into smaller components (or even roots) but also
the identification of instances that should remain unsplitted as
they are of idiomatic nature. We develop a two-fold deep learning-
based approach of noun compound splitting and idiomatic compound
detection for the German language that we train using a newly
collected corpus of annotated German compounds. Our neural noun
compound splitter operates on a sub-word level and outperforms the
current state of the art by about 5\%},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.543}
```

}

@InProceedings{hubers-cucchiarini-strik:2020:LREC,
  author   = {Hubers, Ferdy and Cucchiarini, Catia and Strik, Helmer},
  title   = {Dedicated Language Resources for Interdisciplinary Research on Multiword Expressions: Best Thing since Sliced Bread},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4418--4425},
  abstract = {Multiword expressions such as idioms (beat about the bush), collocations (plastic surgery) and lexical bundles (in the middle of) are challenging for disciplines like Natural Language Processing (NLP), psycholinguistics and second language acquisition, , due to their more or less fixed character. Idiomatic expressions are especially problematic, because they convey a figurative meaning that cannot always be inferred from the literal meanings of the component words. Researchers acknowledge that important properties that characterize idioms such as frequency of exposure, familiarity, transparency, and imageability, should be taken into account in research, but these are typically properties that rely on subjective judgments. This is probably one of the reasons why many studies that investigated idiomatic expressions collected limited information about idiom properties for very small numbers of idioms only. In this paper we report on cross-boundary work aimed at developing a set of tools and language resources that are considered crucial for this kind of multifaceted research. We discuss the results of our research and suggest possible avenues for future research},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.544}
}

@InProceedings{kochmar-gooding-shardlow:2020:LREC,
  author   = {Kochmar, Ekaterina and Gooding, Sian and Shardlow, Matthew},
  title   = {Detecting Multiword Expression Type Helps Lexical Complexity Assessment},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4426--4435},
  abstract = {Multiword expressions (MWEs) represent lexemes that should be treated as single lexical units due to their idiosyncratic nature. Multiple NLP applications have been shown to benefit from MWE identification, however the research on lexical complexity of MWEs is still an under-explored area. In this work, we re-annotate the Complex Word Identification Shared Task 2018 dataset of Yimam et

```
al. (2017), which provides complexity scores for a range of lexemes,
with the types of MWEs. We release the MWE-annotated dataset with
this paper, and we believe this dataset represents a valuable
resource for the text simplification community. In addition, we
investigate which types of expressions are most problematic for
native and non-native readers. Finally, we show that a lexical
complexity assessment system benefits from the information about MWE
types.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.545}
}

@InProceedings{dumitrescu-avram:2020:LREC,
  author    = {Dumitrescu, Stefan Daniel  and  Avram, Andrei-
Marius},
  title     = {Introducing RONEC - the Romanian Named Entity
Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4436--4443},
  abstract  = {We present RONEC - the Named Entity Corpus for the
Romanian language. The corpus contains over 26000 entities in ~5000
annotated sentences, belonging to 16 distinct classes. The sentences
have been extracted from a copy-right free newspaper, covering
several styles. This corpus represents the first initiative in the
Romanian language space specifically targeted for named entity
recognition. It is available in BRAT and CoNLL-U Plus formats, and
it is free to use and extend at github.com/dumitrescustefan/ronec},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.546}
}

@InProceedings{berg-dalianis:2020:LREC,
  author    = {Berg, Hanna  and  Dalianis, Hercules},
  title     = {A Semi-supervised Approach for De-identification of
Swedish Clinical Text},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4444--4450},
  abstract  = {An abundance of electronic health records (EHR) is
produced every day within healthcare. The records possess valuable
information for research and future improvement of healthcare.
Multiple efforts have been done to protect the integrity of patients
while making electronic health records usable for research by
removing personally identifiable information in patient records.
Supervised machine learning approaches for de-identification of EHRs
need annotated data for training, annotations that are costly in
time and human resources. The annotation costs for clinical text is
```

even more costly as the process must be carried out in a protected environment with a limited number of annotators who must have signed confidentiality agreements. In this paper is therefore, a semi-supervised method proposed, for automatically creating high-quality training data. The study shows that the method can be used to improve recall from 84.75\% to 89.20\% without sacrificing precision to the same extent, dropping from 95.73\% to 94.20\%. The model's recall is arguably more important for de-identification than precision.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.547}
}

@InProceedings{lee-EtAl:2020:LREC2,
  author   = {Lee, Chin and Dai, Hongliang and Song, Yangqiu and Li, Xin},
  title   = {A Chinese Corpus for Fine-grained Entity Typing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4451--4457},
  abstract = {Fine-grained entity typing is a challenging task with wide applications. However, most existing datasets for this task are in English. In this paper, we introduce a corpus for Chinese fine-grained entity typing that contains 4,800 mentions manually labeled through crowdsourcing. Each mention is annotated with free-form entity types. To make our dataset useful in more possible scenarios, we also categorize all the fine-grained types into 10 general types. Finally, we conduct experiments with some neural models whose structures are typical in fine-grained entity typing and show how well they perform on our dataset. We also show the possibility of improving Chinese fine-grained entity typing through cross-lingual transfer learning.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.548}
}

@InProceedings{hubkov-kral-pettersson:2020:LREC,
  author   = {Hubková, Helena and Kral, Pavel and Pettersson, Eva},
  title   = {Czech Historical Named Entity Corpus v 1.0},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4458--4465},
  abstract = {As the number of digitized archival documents increases very rapidly, named entity recognition (NER) in historical documents has become very important for information extraction and data mining. For this task an annotated corpus is needed, which has up to now been missing for Czech. In this paper we present a new

```
annotated data collection for historical NER, composed of Czech
historical newspapers. This corpus is freely available for research
purposes. For this corpus, we have defined relevant domain-specific
named entity types and created an annotation manual for corpus
labelling. We further conducted some experiments on this corpus
using recurrent neural networks. We experimented with randomly
initialized embeddings and static and dynamic fastText word
embeddings. We achieved 0.73 F1 score with a bidirectional LSTM
model using static fastText embeddings.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.549}
}

@InProceedings{otegi-EtAl:2020:LREC,
  author    = {Otegi, Arantxa  and  Agirre, Aitor  and  Campos, Jon
Ander  and  Soroa, Aitor  and  Agirre, Eneko},
  title     = {Conversational Question Answering in Low Resource
Scenarios: A Dataset and Case Study for Basque},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {436--442},
  abstract  = {Conversational Question Answering (CQA) systems meet
user information needs by having conversations with them, where
answers to the questions are retrieved from text. There exist a
variety of datasets for English, with tens of thousands of training
examples, and pre-trained language models have allowed to obtain
impressive results. The goal of our research is to test the
performance of CQA systems under low-resource conditions which are
common for most non-English languages: small amounts of native
annotations and other limitations linked to low resource languages,
like lack of crowdworkers or smaller wikipedias. We focus on the
Basque language, and present the first non-English CQA dataset and
results. Our experiments show that it is possible to obtain good
results with low amounts of native data thanks to cross-lingual
transfer, with quality comparable to those obtained for English. We
also discovered that dialogue history models are not directly
transferable to another language, calling for further research. The
dataset is publicly available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.55}
}

@InProceedings{eder-kriegholz-hahn:2020:LREC,
  author    = {Eder, Elisabeth  and  Krieg-Holz, Ulrike  and  Hahn,
Udo},
  title     = {CodE Alltag 2.0 – A Pseudonymized German-Language
Email Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {4466--4477},
  abstract  = {The vast amount of social communication distributed
over various electronic media channels (tweets, blogs, emails,
etc.), so-called user-generated content (UGC), creates entirely new
opportunities for today's NLP research. Yet, data privacy concerns
implied by the unauthorized use of these text streams as a data
resource are often neglected. In an attempt to reconciliate the
diverging needs of unconstrained raw data use and preservation of
data privacy in digital communication, we here investigate the
automatic recognition of privacy-sensitive stretches of text in UGC
and provide an algorithmic solution for the protection of personal
data via pseudonymization. Our focus is directed at the de-
identification of emails where personally identifying information
does not only refer to the sender but also to those people,
locations, dates, and other identifiers mentioned in greetings,
boilerplates and the content-carrying body of emails. We evaluate
several de-identification procedures and systems on two hitherto
non-anonymized German-language email corpora (CodE AlltagS+d and
CodE AlltagXL), and generate fully pseudonymized versions for both
(CodE Alltag 2.0) in which personally identifying information of all
social actors addressed in these mails has been camouflaged (to the
greatest extent possible).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.550}
}

@InProceedings{leitner-rehm-morenoschneider:2020:LREC,
  author    = {Leitner, Elena  and  Rehm, Georg  and  Moreno-
Schneider, Julian},
  title     = {A Dataset of German Legal Documents for Named Entity
Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4478--4485},
  abstract  = {We describe a dataset developed for Named Entity
Recognition in German federal court decisions. It consists of
approx. 67,000 sentences with over 2 million tokens. The resource
contains 54,000 manually annotated entities, mapped to 19 fine-
grained semantic classes: person, judge, lawyer, country, city,
street, landscape, organization, company, institution, court, brand,
law, ordinance, European legal norm, regulation, contract, court
decision, and legal literature. The legal documents were,
furthermore, automatically annotated with more than 35,000 TimeML-
based time expressions. The dataset, which is available under a CC-
BY 4.0 license in the CoNNL-2002 format, was developed for training
an NER service for German legal documents in the EU project Lynx.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.551}
}

@InProceedings{garcapablos-perez-cuadros:2020:LREC,
```

  author   = {García Pablos, Aitor and Perez, Naiara and Cuadros, Montse},
  title   = {Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4486--4494},
  abstract = {Massive digital data processing provides a wide range of opportunities and benefits, but at the cost of endangering personal data privacy. Anonymisation consists in removing or replacing sensitive information from data, enabling its exploitation for different purposes while preserving the privacy of individuals. Over the years, a lot of automatic anonymisation systems have been proposed; however, depending on the type of data, the target language or the availability of training documents, the task remains challenging still. The emergence of novel deep-learning models during the last two years has brought large improvements to the state of the art in the field of Natural Language Processing. These advancements have been most noticeably led by BERT, a model proposed by Google in 2018, and the shared language models pre-trained on millions of documents. In this paper, we use a BERT-based sequence labelling model to conduct a series of anonymisation experiments on several clinical datasets in Spanish. We also compare BERT with other algorithms. The experiments show that a simple BERT-based model with general-domain pre-training obtains highly competitive results without any domain specific feature engineering.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.552}
}

@InProceedings{schulz-EtAl:2020:LREC,
  author   = {Schulz, Sarah and Ševa, Jurica and Rodriguez, Samuel and Ostendorff, Malte and Rehm, Georg},
  title   = {Named Entities in Medical Case Reports: Corpus and Experiments},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4495--4500},
  abstract = {We present a new corpus comprising annotations of medical entities in case reports, originating from PubMed Central’s open access library. In the case reports, we annotate cases, conditions, findings, factors and negation modifiers. Moreover, where applicable, we annotate relations between these entities. As such, this is the first corpus of this kind made available to the scientific community in English. It enables the initial investigation of automatic information extraction from case reports through tasks like Named Entity Recognition, Relation Extraction and

```
(sentence/paragraph) relevance detection. Additionally, we present
four strong baseline systems for the detection of medical entities
made available through the annotated dataset.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.553}
}

@InProceedings{klang-nugues:2020:LREC,
  author    = {Klang, Marcus  and  Nugues, Pierre},
  title     = {Hedwig: A Named Entity Linker},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4501--4508},
  abstract  = {Named entity linking is the task of identifying
mentions of named things in text, such as "Barack Obama" or "New
York", and linking these mentions to unique identifiers. In this
paper, we describe Hedwig, an end-to-end named entity linker, which
uses a combination of word and character BILSTM models for mention
detection, a Wikidata and Wikipedia-derived knowledge base with
global information aggregated over nine language editions, and a
PageRank algorithm for entity linking. We evaluated Hedwig on the
TAC2017 dataset, consisting of news texts and discussion forums, and
we obtained a final score of 59.9\% on CEAFmC+, an improvement over
our previous generation linker Ugglan, and a trilingual entity link
score of 71.9\%.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.554}
}

@InProceedings{barreaux-besagni:2020:LREC,
  author    = {Barreaux, Sabine  and  Besagni, Dominique},
  title     = {An Experiment in Annotating Animal Species Names from
ISTEX Resources},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4509--4513},
  abstract  = {To exploit scientific publications from global
research for TDM purposes, the ISTEX platform enriched its data with
value-added information to ease access to its full-text documents.
We built an experiment to explore new enrichment possibilities in
documents focussing on scientific named entities recognition which
could be integrated into ISTEX resources. This led to testing two
detection tools for animal species names in a corpus of 100
documents in zoology. This makes it possible to provide the French
scientific community with an annotated reference corpus available
for use to measure these tools' performance.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.555}
}
```

```
@InProceedings{caubrire-EtAl:2020:LREC,
  author    = {Caubrière, Antoine  and  Rosset, Sophie  and  Estève, Yannick  and  Laurent, Antoine  and  Morin, Emmanuel},
  title     = {Where are we in Named Entity Recognition from Speech?},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4514--4520},
  abstract  = {Named entity recognition (NER) from speech is usually made through a pipeline process that consists in (i) processing audio using an automatic speech recognition system (ASR) and (ii) applying a NER to the ASR outputs. The latest data available for named entity extraction from speech in French were produced during the ETAPE evaluation campaign in 2012. Since the publication of ETAPE's campaign results, major improvements were done on NER and ASR systems, especially with the development of neural approaches for both of these components. In addition, recent studies have shown the capability of End-to-End (E2E) approach for NER / SLU tasks. In this paper, we propose a study of the improvements made in speech recognition and named entity recognition for pipeline approaches. For this type of systems, we propose an original 3-pass approach. We also explore the capability of an E2E system to do structured NER. Finally, we compare the performances of ETAPE's systems (state-of-the-art systems in 2012) with the performances obtained using current technologies. The results show the interest of the E2E approach, which however remains below an updated pipeline approach.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.556}
}

@InProceedings{mcnamee-EtAl:2020:LREC,
  author    = {McNamee, Paul  and  Mayfield, James  and  Costello, Cash  and  Bishop, Caitlyn  and  Anderson, Shelby},
  title     = {Tagging Location Phrases in Text},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4521--4528},
  abstract  = {For over thirty years researchers have studied the problem of automatically detecting named entities in written language. Throughout this time the majority of such work has focused on detection and classification of entities into coarse-grained types like: PERSON, ORGANIZATION, and LOCATION. Less attention has been focused on non-named mentions of entities, including non-named location phrases such as "the medical clinic in Telonge" or "2 km below the Dolin Maniche bridge". In this work we describe the
```

```
Location Phrase Detection task to identify such spans. Our key
accomplishments include: developing a sequential tagging approach;
crafting annotation guidelines; building annotated datasets for
English and Russian news; and, conducting experiments in automated
detection of location phrases with both statistical and neural
taggers. This work is motivated by extracting rich location
information to support situational awareness during humanitarian
crises such as natural disasters.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.557}
}

@InProceedings{smith-EtAl:2020:LREC,
  author    = {Smith, Hannah  and  Zhang, Zeyu  and  Culnan, John
and  Jansen, Peter},
  title     = {ScienceExamCER: A High-Density Fine-Grained Science-
Domain Corpus for Common Entity Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4529--4546},
  abstract  = {Named entity recognition identifies common classes of
entities in text, but these entity labels are generally sparse,
limiting utility to downstream tasks. In this work we present
ScienceExamCER, a densely-labeled semantic classification corpus of
133k mentions in the science exam domain where nearly all (96\%) of
content words have been annotated with one or more fine-grained
semantic class labels including taxonomic groups, meronym groups,
verb/action groups, properties and values, and synonyms. Semantic
class labels are drawn from a manually-constructed fine-grained
typology of 601 classes generated through a data-driven analysis of
4,239 science exam questions. We show an off-the-shelf BERT-based
named entity recognition model modified for multi-label
classification achieves an accuracy of 0.85 F1 on this task,
suggesting strong utility for downstream tasks in science domain
question answering requiring densely-labeled semantic
classification.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.558}
}

@InProceedings{jrgensen-EtAl:2020:LREC,
  author    = {Jørgensen, Fredrik  and  Aasmoe, Tobias  and  Ruud
Husevåg, Anne-Stine  and  Øvrelid, Lilja  and  Velldal, Erik},
  title     = {NorNE: Annotating Named Entities for Norwegian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4547--4556},
  abstract  = {This paper presents NorNE, a manually annotated
```

corpus of named entities which extends the annotation of the existing Norwegian Dependency Treebank. Comprising both of the official standards of written Norwegian (Bokmål and Nynorsk), the corpus contains around 600,000 tokens and annotates a rich set of entity types including persons, organizations, locations, geo-political entities, products, and events, in addition to a class corresponding to nominals derived from names. We here present details on the annotation effort, guidelines, inter-annotator agreement and an experimental analysis of the corpus using a neural sequence labeling architecture.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.559}
}

@InProceedings{yamazaki-EtAl:2020:LREC,
  author   = {Yamazaki, Yoshihiro and Chiba, Yuya and Nose, Takashi and Ito, Akinori},
  title   = {Construction and Analysis of a Multimodal Chat-talk Corpus for Dialog Systems Considering Interpersonal Closeness},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {443--448},
  abstract = {There are high expectations for multimodal dialog systems that can make natural small talk with facial expressions, gestures, and gaze actions as next-generation dialog-based systems. Two important roles of the chat-talk system are keeping the user engaged and establishing rapport. Many studies have conducted user evaluations of such systems, some of which reported that considering the relationship with the user is an effective way to improve the subjective evaluation. To facilitate research of such dialog systems, we are currently constructing a large-scale multimodal dialog corpus focusing on the relationship between speakers. In this paper, we describe the data collection and annotation process, and analysis of the corpus collected in the early stage of the project. This corpus contains 19,303 utterances (10 hours) from 19 pairs of participants. A dialog act tag is annotated to each utterance by two annotators. We compare the frequency and the transition probability of the tags between different closeness levels to help construct a dialog system for establishing a relationship with the user.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.56}
}

@InProceedings{lffler-EtAl:2020:LREC,
  author   = {Löffler, Felicitas and Abdelmageed, Nora and Babalou, Samira and Kaur, Pawandeep and König-Ries, Birgitta},
  title   = {Tag Me If You Can! Semantic Annotation of Biodiversity Metadata with the QEMP Corpus and the BiodivTagger},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4557--4564},
  abstract  = {Dataset Retrieval is gaining importance due to a
large amount of research data and the great demand for reusing
scientific data. Dataset Retrieval is mostly based on metadata,
structured information about the primary data. Enriching these
metadata with semantic annotations based on Linked Open Data (LOD)
enables datasets, publications and authors to be connected and
expands the search on semantically related terms. In this work, we
introduce the BiodivTagger, an ontology-based Information Extraction
pipeline, developed for metadata from biodiversity research. The
system recognizes biological, physical and chemical processes,
environmental terms, data parameters and phenotypes as well as
materials and chemical compounds and links them to concepts in
dedicated ontologies. To evaluate our pipeline, we created a gold
standard of 50 metadata files (QEMP corpus) selected from five
different data repositories in biodiversity research. To the best of
our knowledge, this is the first annotated metadata corpus for
biodiversity research data. The results reveal a mixed picture.
While materials and data parameters are properly matched to
ontological concepts in most cases, some ontological issues occurred
for processes and environmental terms.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.560}
}

@InProceedings{yada-EtAl:2020:LREC,
  author    = {Yada, Shuntaro  and  Joh, Ayami  and  Tanaka, Ribeka
and  Cheng, Fei  and  Aramaki, Eiji  and  Kurohashi, Sadao},
  title     = {Towards a Versatile Medical-Annotation Guideline
Feasible Without Heavy Medical Knowledge: Starting From Critical
Lung Diseases},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4565--4572},
  abstract  = {Applying natural language processing (NLP) to medical
and clinical texts can bring important social benefits by mining
valuable information from unstructured text. A popular application
for that purpose is named entity recognition (NER), but the
annotation policies of existing clinical corpora have not been
standardized across clinical texts of different types. This paper
presents an annotation guideline aimed at covering medical documents
of various types such as radiography interpretation reports and
medical records. Furthermore, the annotation was designed to avoid
burdensome requirements related to medical knowledge, thereby
enabling corpus development without medical specialists. To achieve
these design features, we specifically focus on critical lung
diseases to stabilize linguistic patterns in corpora. After
annotating around 1100 electronic medical records following the
annotation scheme, we demonstrated its feasibility using an NER
```

task. Results suggest that our guideline is applicable to large-scale clinical NLP projects.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.561}
}

@InProceedings{brandsen-EtAl:2020:LREC,
  author   = {Brandsen, Alex and Verberne, Suzan and Wansleeben, Milco and Lambers, Karsten},
  title   = {Creating a Dataset for Named Entity Recognition in the Archaeology Domain},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4573--4577},
  abstract = {In this paper, we present the development of a training dataset for Dutch Named Entity Recognition (NER) in the archaeology domain. This dataset was created as there is a dire need for semantic search within archaeology, in order to allow archaeologists to find structured information in collections of Dutch excavation reports, currently totalling around 60,000 (658 million words) and growing rapidly. To guide this search task, NER is needed. We created rigorous annotation guidelines in an iterative process, then instructed five archaeology students to annotate a number of documents. The resulting dataset contains ~31k annotations between six entity types (artefact, time period, place, context, species \& material). The inter-annotator agreement is 0.95, and when we used this data for machine learning, we observed an increase in F1 score from 0.51 to 0.70 in comparison to a machine learning model trained on a dataset created in prior work. This indicates that the data is of high quality, and can confidently be used to train NER classifiers.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.562}
}

@InProceedings{zhang-EtAl:2020:LREC2,
  author   = {Zhang, Hongkuan and Sasano, Ryohei and Takeda, Koichi and Shui-Yee Wong, Zoie},
  title   = {Development of a Medical Incident Report Corpus with Intention and Factuality Annotation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4578--4584},
  abstract = {Medical incident reports (MIRs) are documents that record what happened in a medical incident. A typical MIR consists of two sections: a structured categorical part and an unstructured text part. Most texts in MIRs describe what medication was intended to be given and what was actually given, because what happened in an

incident is largely due to discrepancies between intended and actual medications. Recognizing the intention of clinicians and the factuality of medication is essential to understand the causes of medical incidents and avoid similar incidents in the future. Therefore, we are developing an MIR corpus with annotation of intention and factuality as well as of medication entities and their relations. In this paper, we present our annotation scheme with respect to the definition of medication entities that we take into account, the method to annotate the relations between entities, and the details of the intention and factuality annotation. We then report the annotated corpus consisting of 349 Japanese medical incident reports.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.563}
}

@InProceedings{faessler-EtAl:2020:LREC,
  author   = {Faessler, Erik and Modersohn, Luise and Lohr, Christina and Hahn, Udo},
  title   = {ProGene - A Large-scale, High-Quality Protein-Gene Annotated Benchmark Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4585--4596},
  abstract = {Genes and proteins constitute the fundamental entities of molecular genetics. We here introduce ProGene (formerly called FSU-PRGE), a corpus that reflects our efforts to cope with this important class of named entities within the framework of a long-lasting large-scale annotation campaign at the Jena University Language \& Information Engineering (JULIE) Lab. We assembled the entire corpus from 11 subcorpora covering various biological domains to achieve an overall subdomain-independent corpus. It consists of 3,308 MEDLINE abstracts with over 36k sentences and more than 960k tokens annotated with nearly 60k named entity mentions. Two annotators strove for carefully assigning entity mentions to classes of genes/proteins as well as families/groups, complexes, variants and enumerations of those where genes and proteins are represented by a single class. The main purpose of the corpus is to provide a large body of consistent and reliable annotations for supervised training and evaluation of machine learning algorithms in this relevant domain. Furthermore, we provide an evaluation of two state-of-the-art baseline systems - BioBert and flair - on the ProGene corpus. We make the evaluation datasets and the trained models available to encourage comparable evaluations of new methods in the future.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.564}
}

@InProceedings{hvingelby-EtAl:2020:LREC,
  author   = {Hvingelby, Rasmus and Pauli, Amalie Brogaard and Barrett, Maria and Rosted, Christina and Lidegaard, Lasse Malm

```
and  Søgaard, Anders},
  title     = {DaNE: A Named Entity Resource for Danish},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4597--4604},
  abstract  = {We present a named entity annotation for the Danish
Universal Dependencies treebank using the CoNLL-2003 annotation
scheme: DaNE. It is the largest publicly available, Danish named
entity gold annotation. We evaluate the quality of our annotations
intrinsically by double annotating the entire treebank and
extrinsically by comparing our annotations to a recently released
named entity annotation of the validation and test sections of the
Danish Universal Dependencies treebank. We benchmark the new
resource by training and evaluating competitive architectures for
supervised named entity recognition (NER), including FLAIR,
monolingual (Danish) BERT and multilingual BERT. We explore cross-
lingual transfer in multilingual BERT from five related languages in
zero-shot and direct transfer setups, and we show that even with our
modestly-sized training set, we improve Danish NER over a recent
cross-lingual approach, as well as over zero-shot transfer from five
related languages. Using multilingual BERT, we achieve higher
performance by fine-tuning on both DaNE and a larger Bokmål
(Norwegian) training set compared to only using DaNE. However, the
highest performance isachieved by using a Danish BERT fine-tuned on
DaNE. Our dataset enables improvements and applicability for Danish
NER beyond cross-lingual methods. We employ a thorough error
analysis of the predictions of the best models for seen and unseen
entities, as well as their robustness on un-capitalized text. The
annotated dataset and all the trained models are made publicly
available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.565}
}

@InProceedings{ruppenhofer-rehbein-flinz:2020:LREC,
  author    = {Ruppenhofer, Josef  and  Rehbein, Ines  and  Flinz,
Carolina},
  title     = {Fine-grained Named Entity Annotations for German
Biographic Interviews},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4605--4614},
  abstract  = {We present a fine-grained NER annotations with 30
labels and apply it to German data. Building on the OntoNotes 5.0
NER inventory, our scheme is adapted for a corpus of transcripts of
biographic interviews by adding categories for AGE and LAN(guage)
and also features extended numeric and temporal categories. Applying
```

```
the scheme to the spoken data as well as a collection of teaser
tweets from newspaper sites, we can confirm its generality for both
domains, also achieving good inter-annotator agreement. We also show
empirically how our inventory relates to the well-established 4-
category NER inventory by re-annotating a subset of the GermEval
2014 NER coarse-grained dataset with our fine label inventory.
Finally, we use a BERT-based system to establish some baseline
models for NER tagging on our two new datasets. Global results in
in-domain testing are quite high on the two datasets, near what was
achieved for the coarse inventory on the CoNLLL2003 data. Cross-
domain testing produces much lower results due to the severe domain
differences.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.566}
}

@InProceedings{luoma-EtAl:2020:LREC,
  author    = {Luoma, Jouni  and  Oinonen, Miika  and  Pyykönen,
Maria  and  Laippala, Veronika  and  Pyysalo, Sampo},
  title     = {A Broad-coverage Corpus for Finnish Named Entity
Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4615--4624},
  abstract  = {We present a new manually annotated corpus for broad-
coverage named entity recognition for Finnish. Building on the
original Universal Dependencies Finnish corpus of 754 documents
(200,000 tokens) representing ten different genres of text, we
introduce annotation marking person, organization, location, product
and event names as well as dates. The new annotation identifies in
total over 10,000 mentions. An evaluation of inter-annotator
agreement indicates that the quality and consistency of annotation
are high, at 94.5\% F-score for exact match. A comprehensive
evaluation using state-of-the-art machine learning methods
demonstrates that the new resource maintains compatibility with a
previously released single-domain corpus for Finnish NER and makes
it possible to recognize named entity mentions in texts drawn from
most domains at precision and recall approaching or exceeding 90\%.
Remaining challenges such as the identification of names in blog
posts and transcribed speech are also identified. The newly
introduced Turku NER corpus and related resources introduced in this
work are released under open licenses via https://turkunlp.org/
turku-ner-corpus .},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.567}
}

@InProceedings{consoli-EtAl:2020:LREC,
  author    = {Consoli, Bernardo  and  Santos, Joaquim  and  Gomes,
Diogo  and  Cordeiro, Fabio  and  Vieira, Renata  and  Moreira,
Viviane},
  title     = {Embeddings for Named Entity Recognition in Geoscience
```

Portuguese Literature},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4625--4630},
  abstract = {This work focuses on Portuguese Named Entity Recognition (NER) in the Geology domain. The only domain-specific dataset in the Portuguese language annotated for NER is the GeoCorpus. Our approach relies on BiLSTM-CRF neural networks (a widely used type of network for this area of research) that use vector and tensor embedding representations. Three types of embedding models were used (Word Embeddings, Flair Embeddings, and Stacked Embeddings) under two versions (domain-specific and generalized). The domain specific Flair Embeddings model was originally trained with a generalized context in mind, but was then fine-tuned with domain-specific Oil and Gas corpora, as there simply was not enough domain corpora to properly train such a model. Each of these embeddings was evaluated separately, as well as stacked with another embedding. Finally, we achieved state-of-the-art results for this domain with one of our embeddings, and we performed an error analysis on the language model that achieved the best results. Furthermore, we investigated the effects of domain-specific versus generalized embeddings.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.568}
}

@InProceedings{ortizsurez-EtAl:2020:LREC,
  author   = {Ortiz Suárez, Pedro Javier and Dupont, Yoann and Muller, Benjamin and Romary, Laurent and Sagot, Benoît},
  title   = {Establishing a New State-of-the-Art for French Named Entity Recognition},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4631--4638},
  abstract = {The French TreeBank developed at the University Paris 7 is the main source of morphosyntactic and syntactic annotations for French. However, it does not include explicit information related to named entities, which are among the most useful information for several natural language processing tasks and applications. Moreover, no large-scale French corpus with named entity annotations contain referential information, which complement the type and the span of each mention with an indication of the entity it refers to. We have manually annotated the French TreeBank with such information, after an automatic pre-annotation step. We sketch the underlying annotation guidelines and we provide a few figures about the resulting annotations.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.569}

}

@InProceedings{vanwaterschoot-EtAl:2020:LREC,
  author   = {van Waterschoot, Jelte and Hendrickx, Iris and Khan, Arif and Klabbers, Esther and de Korte, Marcel and Strik, Helmer and Cucchiarini, Catia and Theune, Mariët},
  title   = {BLISS: An Agent for Collecting Spoken Dialogue Data about Health and Well-being},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {449--458},
  abstract = {An important objective in health-technology is the ability to gather information about people's well-being. Structured interviews can be used to obtain this information, but are time-consuming and not scalable. Questionnaires provide an alternative way to extract such information, though typically lack depth. In this paper, we present our first prototype of the BLISS agent, an artificial intelligent agent which intends to automatically discover what makes people happy and healthy. The goal of Behaviour-based Language-Interactive Speaking Systems (BLISS) is to understand the motivations behind people's happiness by conducting a personalized spoken dialogue based on a happiness model. We built our first prototype of the model to collect 55 spoken dialogues, in which the BLISS agent asked questions to users about their happiness and well-being. Apart from a description of the BLISS architecture, we also provide details about our dataset, which contains over 120 activities and 100 motivations and is made available for usage.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.57}
}

@InProceedings{lawrie-mayfield-etter:2020:LREC,
  author   = {Lawrie, Dawn and Mayfield, James and Etter, David},
  title   = {Building OCR/NER Test Collections},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4639--4646},
  abstract = {Named entity recognition (NER) identifies spans of text that contain names. Many researchers have reported the results of NER on text created through optical character recognition (OCR) over the past two decades. Unfortunately, the test collections that support this research are annotated with named entities after optical character recognition (OCR) has been run. This means that the collection must be re-annotated if the OCR output changes. Instead by tying annotations to character locations on the page, a collection can be built that supports OCR and NER research without

requiring re-annotation when either improves. This means that named entities are annotated on the transcribed text. The transcribed text is all that is needed to evaluate the performance of OCR. For NER evaluation, the tagged OCR output is aligned to the transcriptions the aligned files, creating modified files of each, which are scored. This paper presents a methodology for building such a test collection and releases a collection of Chinese OCR-NER data constructed using the methodology. The paper provides performance baselines for current OCR and NER systems applied to this new collection.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.570}
}

@InProceedings{marinova-EtAl:2020:LREC,
  author   = {Marinova, Iva and Laskova, Laska and Osenova, Petya and Simov, Kiril and Popov, Alexander},
  title   = {Reconstructing NER Corpora: a Case Study on Bulgarian},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4647--4652},
  abstract = {The paper reports on the usage of deep learning methods for improving a Named Entity Recognition (NER) training corpus and for predicting and annotating new types in a test corpus. We show how the annotations in a type-based corpus of named entities (NE) were populated as occurrences within it, thus ensuring density of the training information. A deep learning model was adopted for discovering inconsistencies in the initial annotation and for learning new NE types. The evaluation results get improved after data curation, randomization and deduplication.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.571}
}

@InProceedings{klimt-EtAl:2020:LREC,
  author   = {Klimt, Kira and Braun, Daniel and Schneider, Daniela and Matthes, Florian},
  title   = {MucLex: A German Lexicon for Surface Realisation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4653--4657},
  abstract = {Language resources for languages other than English are often scarce. Rule-based surface realisers need elaborate lexica in order to be able to generate correct language, especially in languages like German, which include many irregular word forms. In this paper, we present MucLex, a German lexicon for the Natural Language Generation task of surface realisation, based on the crowd-

sourced online lexicon Wiktionary. MucLex contains more than 100,000 lemmata and more than 670,000 different word forms in a well-structured XML file and is available under the Creative Commons BY-SA 3.0 license.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.572}
}

@InProceedings{hu-sun:2020:LREC,
  author   = {Hu, Jinyi and Sun, Maosong},
  title   = {Generating Major Types of Chinese Classical Poetry in a Uniformed Framework},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4658--4663},
  abstract = {Poetry generation is an interesting research topic in the field of text generation. As one of the most valuable literary and cultural heritages of China, Chinese classical poetry is very familiar and loved by Chinese people from generation to generation. It has many particular characteristics in its language structure, ranging from form, sound to meaning, thus is regarded as an ideal testing task for text generation. In this paper, we propose a GPT-2 based uniformed framework for generating major types of Chinese classical poems. We define a unified format for formulating all types of training samples by integrating detailed form information, then present a simple form- stressed weighting method in GPT-2 to strengthen the control to the form of the generated poems, with special emphasis on those forms with longer body length. Preliminary experimental results show this enhanced model can generate Chinese classical poems of major types with high quality in both form and content, validating the effectiveness of the proposed strategy. The model has been incorporated into Jiuge, the most influential Chinese classical poetry generation system developed by Tsinghua University.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.573}
}

@InProceedings{shigeto-EtAl:2020:LREC,
  author   = {Shigeto, Yutaro and Yoshikawa, Yuya and Lin, Jiaqing and Takeuchi, Akikazu},
  title   = {Video Caption Dataset for Describing Human Actions in Japanese},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4664--4670},
  abstract = {In recent years, automatic video caption generation has attracted considerable attention. This paper focuses on the

```
generation of Japanese captions for describing human actions. While most currently available video caption datasets have been constructed for English, there is no equivalent Japanese dataset. To address this, we constructed a large-scale Japanese video caption dataset consisting of 79,822 videos and 399,233 captions. Each caption in our dataset describes a video in the form of “who does what and where.” To describe human actions, it is important to identify the details of a person, place, and action. Indeed, when we describe human actions, we usually mention the scene, person, and action. In our experiments, we evaluated two caption generation methods to obtain benchmark results. Further, we investigated whether those generation methods could specify “who does what and where.”},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.574}
}

@InProceedings{wen-EtAl:2020:LREC,
  author    = {Wen, Zhiyuan  and  Cao, Jiannong  and  Yang, Ruosong  and  Wang, Senzhang},
  title     = {Decode with Template: Content Preserving Sentiment Transfer},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4671--4679},
  abstract  = {Sentiment transfer aims to change the underlying sentiment of input sentences. The two major challenges in existing works lie in (1) effectively disentangling the original sentiment from input sentences; and (2) preserving the semantic content while transferring the sentiment. We find that identifying the sentiment-irrelevant content from input sentences to facilitate generating output sentences could address the above challenges and then propose the Decode with Template model in this paper. We first mask the explicit sentiment words in input sentences and use the rest parts as templates to eliminate the original sentiment. Then, we input the templates and the target sentiments into our bidirectionally guided variational auto-encoder (VAE) model to generate output. In our method, the template preserves most of the semantics in input sentences, and the bidirectionally guided decoding captures both forward and backward contextual information to generate output. Both two parts contribute to better content preservation. We evaluate our method on two review datasets, Amazon and Yelp, with automatic evaluation methods and human rating. The experimental results show that our method significantly outperforms state-of-the-art models, especially in content preservation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.575}
}

@InProceedings{sauder-EtAl:2020:LREC,
  author    = {Sauder, Jonathan  and  Hu, Ting  and  Che, Xiaoyin  and  Mordido, Goncalo  and  Yang, Haojin  and  Meinel, Christoph},
```

```
  title     = {Best Student Forcing: A Simple Training Mechanism in
Adversarial Language Generation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4680--4688},
  abstract  = {Language models trained with Maximum Likelihood
Estimation (MLE) have been considered as a mainstream solution in
Natural Language Generation (NLG) for years. Recently, various
approaches with Generative Adversarial Nets (GANs) have also been
proposed. While offering exciting new prospects, GANs in NLG by far
are nevertheless reportedly suffering from training instability and
mode collapse, and therefore outperformed by conventional MLE
models. In this work, we propose techniques for improving GANs in
NLG, namely Best Student Forcing (BSF), a novel yet simple
adversarial training mechanism in which generated sequences of high
quality are selected as temporary ground-truth to further train the
generator. We also use an ensemble of discriminators to increase
training stability and sample diversity. Evaluation shows that the
combination of BSF and multiple discriminators consistently performs
better than previous GAN approaches over various metrics, and
outperforms a baseline MLE in terms of Fr´ech´et Distance, a
recently proposed metric capturing both sample quality and
diversity.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.576}
}

@InProceedings{martin-EtAl:2020:LREC1,
  author    = {Martin, Louis  and  de la Clergerie, Éric  and
Sagot, Benoît  and  Bordes, Antoine},
  title     = {Controllable Sentence Simplification},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4689--4698},
  abstract  = {Text simplification aims at making a text easier to
read and understand by simplifying grammar and structure while
keeping the underlying information identical. It is often considered
an all-purpose generic task where the same simplification is
suitable for all; however multiple audiences can benefit from
simplified text in different ways. We adapt a discrete
parametrization mechanism that provides explicit control on
simplification systems based on Sequence-to-Sequence models. As a
result, users can condition the simplifications returned by a model
on attributes such as length, amount of paraphrasing, lexical
complexity and syntactic complexity. We also show that carefully
chosen values of these attributes allow out-of-the-box Sequence-to-
Sequence models to outperform their standard counterparts on
```

```
simplification benchmarks. Our model, which we call ACCESS (as
shorthand for AudienCe-CEntric Sentence Simplification), establishes
the state of the art at 41.87 SARI on the WikiLarge test set, a
+1.42 improvement over the best previously reported score.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.577}
}

@InProceedings{aminnejad-ive-velupillai:2020:LREC,
  author    = {Amin-Nejad, Ali  and  Ive, Julia  and  Velupillai,
Sumithra},
  title     = {Exploring Transformer Text Generation for Medical
Dataset Augmentation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4699--4708},
  abstract  = {Natural Language Processing (NLP) can help unlock the
vast troves of unstructured data in clinical text and thus improve
healthcare research. However, a big barrier to developments in this
field is data access due to patient confidentiality which prohibits
the sharing of this data, resulting in small, fragmented and
sequestered openly available datasets. Since NLP model development
requires large quantities of data, we aim to help side-step this
roadblock by exploring the usage of Natural Language Generation in
augmenting datasets such that they can be used for NLP model
development on downstream clinically relevant tasks. We propose a
methodology guiding the generation with structured patient
information in a sequence-to-sequence manner. We experiment with
state-of-the-art Transformer models and demonstrate that our
augmented dataset is capable of beating our baselines on a
downstream classification task. Finally, we also create a user
interface and release the scripts to train generation models to
stimulate further research in this area.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.578}
}

@InProceedings{liyanage-ranathunga:2020:LREC,
  author    = {Liyanage, Vijini  and  Ranathunga, Surangika},
  title     = {Multi-lingual Mathematical Word Problem Generation
using Long Short Term Memory Networks with Enhanced Input Features},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4709--4716},
  abstract  = {A Mathematical Word Problem (MWP) differs from a
general textual representation due to the fact that it is comprised
of numerical quantities and units, in addition to text. Therefore,
MWP generation should be carefully handled. When it comes to multi-
```

lingual MWP generation, language specific morphological and syntactic features become additional constraints. Standard template-based MWP generation techniques are incapable of identifying these language specific constraints, particularly in morphologically rich yet low resource languages such as Sinhala and Tamil. This paper presents the use of a Long Short Term Memory (LSTM) network that is capable of generating elementary level MWPs, while satisfying the aforementioned constraints. Our approach feeds a combination of character embeddings, word embeddings, and Part of Speech (POS) tag embeddings to the LSTM, in which attention is provided for numerical values and units. We trained our model for three languages, English, Sinhala and Tamil using separate MWP datasets. Irrespective of the language and the type of the MWP, our model could generate accurate single sentenced and multi sentenced problems. Accuracy reported in terms of average BLEU score for English, Sinhala and Tamil languages were 22.97\%, 24.49\% and 20.74\%, respectively.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.579}
}

@InProceedings{chen-EtAl:2020:LREC1,
  author   = {Chen, Meng and Liu, Ruixue and Shen, Lei and Yuan, Shaozu and Zhou, Jingyan and Wu, Youzheng and He, Xiaodong and Zhou, Bowen},
  title   = {The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {459--466},
  abstract = {Human conversations are complicated and building a human-like dialogue agent is an extremely challenging task. With the rapid development of deep learning techniques, data-driven models become more and more prevalent which need a huge amount of real conversation data. In this paper, we construct a large-scale real scenario Chinese E-commerce conversation corpus, JDDC, with more than 1 million multi-turn dialogues, 20 million utterances, and 150 million words. The dataset reflects several characteristics of human-human conversations, e.g., goal-driven, and long-term dependency among the context. It also covers various dialogue types including task-oriented, chitchat and question-answering. Extra intent information and three well-annotated challenge sets are also provided. Then, we evaluate several retrieval-based and generative models to provide basic benchmark performance on the JDDC corpus. And we hope JDDC can serve as an effective testbed and benefit the development of fundamental research in dialogue task.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.58}
}

@InProceedings{doughman-abusalem-elbassuoni:2020:LREC,
  author   = {Doughman, Jad and Abu Salem, Fatima and Elbassuoni, Shady},

```
  title     = {Time-Aware Word Embeddings for Three Lebanese News
Archives},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4717--4725},
  abstract  = {Word embeddings have proven to be an effective method
for capturing semantic relations among distinct terms within a large
corpus. In this paper, we present a set of word embeddings learnt
from three large Lebanese news archives, which collectively consist
of 609,386 scanned newspaper images and spanning a total of 151
years, ranging from 1933 till 2011. The diversified ideological
nature of the news archives alongside the temporal variability of
the embeddings offer a rare glimpse onto the variation of word
representation across the left-right political spectrum. To train
the word embeddings, Google's Tesseract 4.0 OCR engine was employed
to transcribe the scanned news archives, and various archive-level
as well as decade-level word embeddings were learnt. To evaluate the
accuracy of the learnt word embeddings, a benchmark of analogy tasks
was used. Finally, we demonstrate an interactive system that allows
the end user to visualize for a given word of interest, the
variation of the top-k closest words in the embedding space as a
function of time and across news archives using an animated scatter
plot.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.580}
}

@InProceedings{yang-cao-wen:2020:LREC,
  author    = {Yang, Ruosong  and  Cao, Jiannong  and  Wen,
Zhiyuan},
  title     = {GGP: Glossary Guided Post-processing for Word
Embedding Learning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4726--4730},
  abstract  = {Word embedding learning is the task to map each word
into a low-dimensional and continuous vector based on a large
corpus. To enhance corpus based word embedding models, researchers
utilize domain knowledge to learn more distinguishable
representations via joint optimization and post-processing based
models. However, joint optimization based models require much
training time. Existing post-processing models mostly consider
semantic knowledge while learned embedding models show less
functional information. Glossary is a comprehensive linguistic
resource. And in previous works, the glossary is usually used to
enhance the word representations via joint optimization based
methods. In this paper, we post-process pre-trained word embedding
```

```
models with incorporating the glossary and capture more topical and
functional information. We propose GGP (Glossary Guided Post-
processing word embedding) model which consists of a global post-
processing function to fine-tune each word vector, and an auto-
encoding model to learn sense representations, furthermore,
constrains each post-processed word representation and the
composition of its sense representations to be similar. We evaluate
our model by comparing it with two state-of-the-art models on six
word topical/functional similarity datasets, and the results show
that it outperforms competitors by an average of 4.1\% across all
datasets. And our model outperforms GloVe by more than 7\%.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.581}
}

@InProceedings{ular-robnikikonja:2020:LREC,
  author    = {Ulčar, Matej  and  Robnik-Šikonja, Marko},
  title     = {High Quality ELMo Embeddings for Seven Less-Resourced
Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4731--4738},
  abstract  = {Recent results show that deep neural networks using
contextual embeddings significantly outperform non-contextual
embeddings on a majority of text classification task. We offer
precomputed embeddings from popular contextual ELMo model for seven
languages: Croatian, Estonian, Finnish, Latvian, Lithuanian,
Slovenian, and Swedish. We demonstrate that the quality of
embeddings strongly depends on the size of training set and show
that existing publicly available ELMo embeddings for listed
languages shall be improved. We train new ELMo embeddings on much
larger training sets and show their advantage over baseline non-
contextual FastText embeddings. In evaluation, we use two
benchmarks, the analogy task and the NER task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.582}
}

@InProceedings{schneider-EtAl:2020:LREC,
  author    = {Schneider, Rudolf  and  Oberhauser, Tom  and
Grundmann, Paul  and  Gers, Felix Alexander  and  Loeser, Alexander
and  Staab, Steffen},
  title     = {Is Language Modeling Enough? Evaluating Effective
Embedding Combinations},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4739--4748},
  abstract  = {Universal embeddings, such as BERT or ELMo, are
```

```
useful for a broad set of natural language processing tasks like
text classification or sentiment analysis. Moreover, specialized
embeddings also exist for tasks like topic modeling or named entity
disambiguation. We study if we can complement these universal
embeddings with specialized embeddings. We conduct an in-depth
evaluation of nine well known natural language understanding tasks
with SentEval. Also, we extend SentEval with two additional tasks to
the medical domain. We present PubMedSection, a novel topic
classification dataset focussed on the biomedical domain. Our
comprehensive analysis covers 11 tasks and combinations of six
embeddings. We report that combined embeddings outperform state of
the art universal embeddings without any embedding fine-tuning. We
observe that adding topic model based embeddings helps for most
tasks and that differing pre-training tasks encode complementary
features. Moreover, we present new state of the art results on the
MPQA and SUBJ tasks in SentEval.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.583}
}

@InProceedings{maupom-meurs:2020:LREC,
  author    = {Maupomé, Diego  and  Meurs, Marie-Jean},
  title     = {Language Modeling with a General Second-Order RNN},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4749--4753},
  abstract  = {Different Recurrent Neural Network (RNN)
architectures update their state in different manners as the input
sequence is processed. RNNs including a multiplicative interaction
between their current state and the current input, second-order
ones, show promising performance in language modeling. In this
paper, we introduce a second-order RNNs that generalizes existing
ones. Evaluating on the Penn Treebank dataset, we analyze how its
different components affect its performance in character-lever
recurrent language modeling. We perform our experiments controlling
the parameter counts of models. We find that removing the first-
order terms does not hinder performance. We perform further
experiments comparing the effects of the relative size of the state
space and the multiplicative interaction space on performance. Our
expectation was that a larger states would benefit language models
built on longer documents, and larger multiplicative interaction
states would benefit ones built on larger input spaces. However, our
results suggest that this is not the case and the optimal relative
size is the same for both document tokenizations used.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.584}
}

@InProceedings{schneidermann-hvingelby-pedersen:2020:LREC,
  author    = {Schneidermann, Nina  and  Hvingelby, Rasmus  and
Pedersen, Bolette},
  title     = {Towards a Gold Standard for Evaluating Danish Word
```

```
Embeddings},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4754--4763},
  abstract  = {This paper presents the process of compiling a model-
agnostic similarity goal standard for evaluating Danish word
embeddings based on human judgments made by 42 native speakers of
Danish. Word embeddings resemble semantic similarity solely by
distribution (meaning that word vectors do not reflect relatedness
as differing from similarity), and we argue that this generalization
poses a problem in most intrinsic evaluation scenarios. In order to
be able to evaluate on both dimensions, our human-generated dataset
is therefore designed to reflect the distinction between relatedness
and similarity. The goal standard is applied for evaluating the
"goodness" of six existing word embedding models for Danish, and it
is discussed how a relatively low correlation can be explained by
the fact that semantic similarity is substantially more challenging
to model than relatedness, and that there seems to be a need for
future human judgments to measure similarity in full context and
along more than a single spectrum.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.585}
}

@InProceedings{wilson-EtAl:2020:LREC,
  author    = {Wilson, Steven  and  Magdy, Walid  and  McGillivray,
Barbara  and  Garimella, Kiran  and  Tyson, Gareth},
  title     = {Urban Dictionary Embeddings for Slang NLP
Applications},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4764--4773},
  abstract  = {The choice of the corpus on which word embeddings are
trained can have a sizable effect on the learned representations,
the types of analyses that can be performed with them, and their
utility as features for machine learning models. To contribute to
the existing sets of pre-trained word embeddings, we introduce and
release the first set of word embeddings trained on the content of
Urban Dictionary, a crowd-sourced dictionary for slang words and
phrases. We show that although these embeddings are trained on fewer
total tokens (by at least an order of magnitude compared to most
popular pre-trained embeddings), they have high performance across a
range of common word embedding evaluations, ranging from semantic
similarity to word clustering tasks. Further, for some extrinsic
tasks such as sentiment analysis and sarcasm detection where we
expect to require some knowledge of colloquial language on social
media data, initializing classifiers with the Urban Dictionary
```

```
Embeddings resulted in improved performance compared to initializing
with a range of other well-known, pre-trained embeddings that are
order of magnitude larger in size.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.586}
}

@InProceedings{kim-kim-lee:2020:LREC,
  author    = {Kim, Yeachan  and  Kim, Kang-Min  and  Lee,
SangKeun},
  title     = {Representation Learning for Unseen Words by Bridging
Subwords to Semantic Networks},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4774--4780},
  abstract  = {Pre-trained word embeddings are widely used in
various fields. However, the coverage of pre-trained word embeddings
only includes words that appeared in corpora where pre-trained
embeddings are learned. It means that the words which do not appear
in training corpus are ignored in tasks, and it could lead to the
limited performance of neural models. In this paper, we propose a
simple yet effective method to represent out-of-vocabulary (OOV)
words. Unlike prior works that solely utilize subword information or
knowledge, our method makes use of both information to represent OOV
words. To this end, we propose two stages of representation
learning. In the first stage, we learn subword embeddings from the
pre-trained word embeddings by using an additive composition
function of subwords. In the second stage, we map the learned
subwords into semantic networks (e.g., WordNet). We then re-train
the subword embeddings by using lexical entries on semantic lexicons
that could include newly observed subwords. This two-stage learning
makes the coverage of words broaden to a great extent. The
experimental results clearly show that our method provides
consistent performance improvements over strong baselines that use
subwords or lexical resources separately.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.587}
}

@InProceedings{agerri-EtAl:2020:LREC,
  author    = {Agerri, Rodrigo  and  San Vicente, Iñaki  and
Campos, Jon Ander  and  Barrena, Ander  and  Saralegi, Xabier  and
Soroa, Aitor  and  Agirre, Eneko},
  title     = {Give your Text Representation Models some Love: the
Case for Basque},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4781--4788},
```

  abstract = {Word embeddings and pre-trained language models allow to build rich representations of text and have enabled improvements across most NLP tasks. Unfortunately they are very expensive to train, and many small companies and research groups tend to use models that have been pre-trained and made available by third parties, rather than building their own. This is suboptimal as, for many languages, the models have been trained on smaller (or lower quality) corpora. In addition, monolingual pre-trained models for non-English languages are not always available. At best, models for those languages are included in multilingual versions, where each language shares the quota of substrings and parameters with the rest of the languages. This is particularly true for smaller languages such as Basque. In this paper we show that a number of monolingual models (FastText word embeddings, FLAIR and BERT language models) trained with larger Basque corpora produce much better results than publicly available versions in downstream NLP tasks, including topic classification, sentiment classification, PoS tagging and NER. This work sets a new state-of-the-art in those tasks for Basque. All benchmarks and models used in this work are publicly available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.588}
}

@InProceedings{torregrossa-EtAl:2020:LREC,
  author   = {Torregrossa, François and Claveau, Vincent and Kooli, Nihel and Gravier, Guillaume and Allesiardo, Robin},
  title   = {On the Correlation of Word Embedding Evaluation Metrics},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4789--4797},
  abstract = {Word embeddings intervene in a wide range of natural language processing tasks. These geometrical representations are easy to manipulate for automatic systems. Therefore, they quickly invaded all areas of language processing. While they surpass all predecessors, it is still not straightforward why and how they do so. In this article, we propose to investigate all kind of evaluation metrics on various datasets in order to discover how they correlate with each other. Those correlations lead to 1) a fast solution to select the best word embeddings among many others, 2) a new criterion that may improve the current state of static Euclidean word embeddings, and 3) a way to create a set of complementary datasets, i.e. each dataset quantifies a different aspect of word embeddings.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.589}
}

@InProceedings{priegovalverde-bigi-amoyal:2020:LREC,
  author   = {Priego-Valverde, Béatrice and Bigi, Brigitte and Amoyal, Mary},
  title   = {"Cheese!": a Corpus of Face-to-face French

Interactions. A Case Study for Analyzing Smiling and Conversational Humor},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {467--475},
  abstract = {Cheese! is a conversational corpus. It consists of 11 French face-to-face conversations lasting around 15 minutes each. Cheese! is a duplication of an American corpus (ref) in order to conduct a cross-cultural comparison of participants' smiling behavior in humorous and non-humorous sequences in American English and French conversations. In this article, the methodology used to collect and enrich the corpus is presented: experimental protocol, technical choices, transcription, semi-automatic annotations, manual annotations of smiling and humor. An exploratory study investigating the links between smile and humor is then proposed. Based on the analysis of two interactions, two questions are asked: (1) Does smile frame humor? (2) Does smile has an impact on its success or failure? If the experimental design of Cheese! has been elaborated to study specifically smiles and humor in conversations, the high quality of the dataset obtained, and the methodology used are also replicable and can be applied to analyze many other conversational activities and other multimodal modalities.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.59}
}

@InProceedings{novk-laki-novk:2020:LREC,
  author   = {Novák, Attila and Laki, László and Novák, Borbála},
  title   = {CBOW-tag: a Modified CBOW Algorithm for Generating Embedding Models from Annotated Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4798--4801},
  abstract = {In this paper, we present a modified version of the CBOW algorithm implemented in the fastText framework. Our modified algorithm, CBOW-tag builds a vector space model that includes the representation of the original word forms and their annotation at the same time. We illustrate the results by presenting a model built from a corpus that includes morphological and syntactic annotations. The simultaneous presence of unannotated elements and different annotations at the same time in the model makes it possible to constrain nearest neighbour queries to specific types of elements. The model can thus efficiently answer questions such as What do we eat?, What can we do with a skeleton? What else do we do with what we eat?, etc. Error analysis reveals that the model can highlight errors introduced into the annotation by the tagger and parser we

```
used to generate the annotations as well as lexical peculiarities in
the corpus itself, especially if we do not limit the vocabulary of
the model to frequent items.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.590}
}

@InProceedings{dmtr-yang-novk:2020:LREC,
  author    = {Dömötör, Andrea  and  Yang, Zijian Győző  and  Novák,
Attila},
  title     = {Much Ado About Nothing – Identification of Zero
Copulas in Hungarian Using an NMT Model},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4802--4810},
  abstract  = {The research presented in this paper concerns zero
copulas in Hungarian, i.e. the phenomenon that nominal predicates
lack an explicit verbal copula in the default present tense 3rd
person indicative case. We created a tool based on the state-of-the-
art transformer architecture implemented in Marian NMT framework
that can identify and mark the location of zero copulas, i.e. the
position where an overt copula would appear in the non-default
cases. Our primary aim was to support quantitative corpus-based
linguistic research by creating a tool that can be used to compile a
corpus of significant size containing examples of nominal predicates
including the location of the zero copulas. We created the training
corpus for our system transforming sentences containing overt
copulas into ones containing zero copula labels. However, we first
needed to disambiguate occurrences of the massively ambiguous verb
van `exist/be/have'. We performed this using a rule-base classifier
relying on English translations in the English-Hungarian parallel
subcorpus of the OpenSubtitles corpus. We created several NMT-based
models using different sampling methods and optionally using our
baseline model to synthesize additional training data. Our best
model obtains almost 90\% precision and 80\% recall on an in-domain
test set.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.591}
}

@InProceedings{martinc-kraljnovak-pollak:2020:LREC,
  author    = {Martinc, Matej  and  Kralj Novak, Petra  and  Pollak,
Senja},
  title     = {Leveraging Contextual Embeddings for Detecting
Diachronic Semantic Shift},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4811--4819},
```

  abstract = {We propose a new method that leverages contextual embeddings for the task of diachronic semantic shift detection by generating time specific word representations from BERT embeddings. The results of our experiments in the domain specific LiverpoolFC corpus suggest that the proposed method has performance comparable to the current state-of-the-art without requiring any time consuming domain adaptation on large corpora. The results on the newly created Brexit news corpus suggest that the method can be successfully used for the detection of a short-term yearly semantic shift. And lastly, the model also shows promising results in a multilingual settings, where the task was to detect differences and similarities between diachronic semantic shifts in different languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.592}
}

@InProceedings{dougal-lonsdale:2020:LREC,
  author   = {Dougal, Duane K. and Lonsdale, Deryle},
  title   = {Improving NMT Quality Using Terminology Injection},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4820--4827},
  abstract = {Many organizations use domain- or organization-specific words and phrases. This paper explores the use of vetted terminology as an input to neural machine translation (NMT) for improved results: ensuring that the translation of individual terms is consistent with an approved multilingual terminology collection. We discuss, implement, and evaluate a method for injecting terminology and for evaluating terminology injection. Our use of the long short-term memory (LSTM) attention mechanism prevalent in state-of-the-art NMT systems involves attention vectors for correctly identifying semantic entities and aligning the tokens that represent them, both in the source and the target languages. Appropriate terminology is then injected into matching alignments during decoding. We also introduce a new translation metric more sensitive to approved terminological content in MT output.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.593}
}

@InProceedings{santos-consoli-vieira:2020:LREC,
  author   = {Santos, Joaquim and Consoli, Bernardo and Vieira, Renata},
  title   = {Word Embedding Evaluation in Downstream Tasks and Semantic Analogies},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4828--4834},

```
  abstract  = {Language Models have long been a prolific area of
study in the field of Natural Language Processing (NLP). One of the
newer kinds of language models, and some of the most used, are Word
Embeddings (WE). WE are vector space representations of a vocabulary
learned by a non-supervised neural network based on the context in
which words appear. WE have been widely used in downstream tasks in
many areas of study in NLP. These areas usually use these vector
models as a feature in the processing of textual data. This paper
presents the evaluation of newly released WE models for the
Portuguese langauage, trained with a corpus composed of 4.9 billion
tokens. The first evaluation presented an intrinsic task in which
WEs had to correctly build semantic and syntactic relations. The
second evaluation presented an extrinsic task in which the WE models
were used in two downstream tasks: Named Entity Recognition and
Semantic Similarity between Sentences. Our results show that a
diverse and comprehensive corpus can often outperform a larger, less
textually diverse corpus, and that batch training may cause quality
loss in WE models.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.594}
}

@InProceedings{lendvai-EtAl:2020:LREC,
  author    = {Lendvai, Piroska  and  Darányi, Sándor  and  Geng,
Christian  and  Kuijpers, Moniek  and  Lopez de Lacalle, Oier  and
Mensonides, Jean-Christophe  and  Rebora, Simone  and  Reichel,
Uwe},
  title     = {Detection of Reading Absorption in User-Generated
Book Reviews: Resources Creation and Evaluation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4835--4841},
  abstract  = {To detect how and when readers are experiencing
engagement with a literary work, we bring together empirical
literary studies and language technology via focusing on the
affective state of absorption. The goal of our resource development
is to enable the detection of different levels of reading absorption
in millions of user-generated reviews hosted on social reading
platforms. We present a corpus of social book reviews in English
that we annotated with reading absorption categories. Based on these
data, we performed supervised, sentence level, binary classification
of the explicit presence vs. absence of the mental state of
absorption. We compared the performances of classical machine
learners where features comprised sentence representations obtained
from a pretrained embedding model (Universal Sentence Encoder) vs.
neural classifiers in which sentence embedding vector
representations are adapted or fine-tuned while training for the
absorption recognition task. We discuss the challenges in creating
the labeled data as well as the possibilities for releasing a
benchmark corpus.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.595}
```

}

@InProceedings{alsudias-rayson:2020:LREC,
  author   = {Alsudias, Lama and Rayson, Paul},
  title   = {Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4842--4850},
  abstract = {Building ontologies is a crucial part of the semantic web endeavour. In recent years, research interest has grown rapidly in supporting languages such as Arabic in NLP in general but there has been very little research on medical ontologies for Arabic. We present a new Arabic ontology in the infectious disease domain to support various important applications including the monitoring of infectious disease spread via social media. This ontology meaningfully integrates the scientific vocabularies of infectious diseases with their informal equivalents. We use ontology learning strategies with manual checking to build the ontology. We applied three statistical methods for term extraction from selected Arabic infectious diseases articles: TF-IDF, C-value, and YAKE. We also conducted a study, by consulting around 100 individuals, to discover the informal terms related to infectious diseases in Arabic. In future work, we will automatically extract the relations for infectious disease concepts but for now these are manually created. We report two complementary experiments to evaluate the ontology. First, a quantitative evaluation of the term extraction results and an additional qualitative evaluation by a domain expert.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.596}
}

@InProceedings{oliver:2020:LREC,
  author   = {Oliver, Antoni},
  title   = {Aligning Wikipedia with WordNet:a Review and Evaluation of Different Techniques},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4851--4858},
  abstract = {In this paper we explore techniques for aligning Wikipedia articles with WordNet synsets, their successful alignment being our main goal. We evaluate techniques that use the definitions and sense relations in Wordnet and the text and categories in Wikipedia articles. The results we present are based on two evaluation strategies: one uses a new gold and silver standard (for which the creation process is explained); the other creates wordnets in other languages and then compares them with existing wordnets for

those languages found in the Open Multilingual Wordnet project. A reliable alignment between WordNet and Wikipedia is a very valuable resource for the creation of new wordnets in other languages and for the development of existing wordnets. The evaluation of alignments between WordNet and lexical resources is a difficult and time-consuming task, but the evaluation strategy using the Open Multilingual Wordnet can be used as an automated evaluation measure to assess the quality of alignments between these two resources.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.597}
}

@InProceedings{branco-EtAl:2020:LREC1,
  author   = {Branco, António and Grilo, Sara and Bolrinha, Márcia and Saedi, Chakaveh and Branco, Ruben and Silva, João and Querido, Andreia and de Carvalho, Rita and Gaudio, Rosa and Avelãs, Mariana and Pinto, Clara},
  title   = {The MWN.PT WordNet for Portuguese: Projection, Validation, Cross-lingual Alignment and Distribution},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4859--4866},
  abstract = {The objective of the present paper is twofold, to present the MWN.PT WordNet and to report on its construction and on the lessons learned with it. The MWN.PT WordNet for Portuguese includes 41,000 concepts, expressed by 38,000 lexical units. Its synsets were manually validated and are linked to semantically equivalent synsets of the Princeton WordNet of English, and thus transitively to the many wordnets for other languages that are also linked to this English wordnet. To the best of our knowledge, it is the largest high quality, manually validated and cross-lingually integrated, wordnet of Portuguese distributed for reuse. Its construction was initiated more than one decade ago and its description is published for the first time in the present paper. It follows a three step <projection, validation with alignment, completion> methodology consisting on the manual validation and expansion of the outcome of an automatic projection procedure of synsets and their hypernym relations, followed by another automatic procedure that transferred the relations of remaining semantic types across wordnets of different languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.598}
}

@InProceedings{bou-EtAl:2020:LREC,
  author   = {Bou, Savong and Suzuki, Naoki and Miwa, Makoto and Sasaki, Yutaka},
  title   = {Ontology-Style Relation Annotation: A Case Study},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4867--4876},
  abstract  = {This paper proposes an Ontology-Style Relation (OSR)
annotation approach. In conventional Relation Extraction (RE)
datasets, relations are annotated as links between entity mentions.
In contrast, in our OSR annotation, a relation is annotated as a
relation mention (i.e., not a link but a node) and domain and range
links are annotated from the relation mention to its argument entity
mentions. We expect the following benefits: (1) the relation
annotations can be easily converted to Resource Description
Framework (RDF) triples to populate an Ontology, (2) some part of
conventional RE tasks can be tackled as Named Entity Recognition
(NER) tasks. The relation classes are limited to several RDF
properties such as domain, range, and subClassOf, and (3) OSR
annotations can be clear documentations of Ontology contents. As a
case study, we converted an in-house corpus of Japanese traffic
rules in conventional annotations into the OSR annotations and built
a novel OSR-RoR (Rules of the Road) corpus. The inter-annotator
agreements of the conversion were 85-87\%. We evaluated the
performance of neural NER and RE tools on the conventional and OSR
annotations. The experimental results showed that the OSR
annotations make the RE task easier while introducing slight
complexity into the NER task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.599}
}

@InProceedings{bamman-lewke-mansoor:2020:LREC,
  author    = {Bamman, David  and  Lewke, Olivia  and  Mansoor,
Anya},
  title     = {An Annotated Dataset of Coreference in English
Literature},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {44--54},
  abstract  = {We present in this work a new dataset of coreference
annotations for works of literature in English, covering 29,103
mentions in 210,532 tokens from 100 works of fiction published
between 1719 and 1922. This dataset differs from previous
coreference corpora in containing documents whose average length
(2,105.3 words) is four times longer than other benchmark datasets
(463.7 for OntoNotes), and contains examples of difficult
coreference problems common in literature. This dataset allows for
an evaluation of cross-domain performance for the task of
coreference resolution, and analysis into the characteristics of
long-distance within-document coreference.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.6}
}

@InProceedings{chierici-habash-bicec:2020:LREC,
```

```
  author    = {Chierici, Alberto  and  Habash, Nizar  and  Bicec,
Margarita},
  title     = {The Margarita Dialogue Corpus: A Data Set for Time-
Offset Interactions and Unstructured Dialogue Systems},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {476--484},
  abstract  = {Time-Offset Interaction Applications (TOIAs) are
systems that simulate face-to-face conversations between humans and
digital human avatars recorded in the past. Developing a well-
functioning TOIA involves several research areas: artificial
intelligence, human-computer interaction, natural language
processing, question answering, and dialogue systems. The first
challenges are to define a sensible methodology for data collection
and to create useful data sets for training the system to retrieve
the best answer to a user's question. In this paper, we present
three main contributions: a methodology for creating the knowledge
base for a TOIA, a dialogue corpus, and baselines for single-turn
answer retrieval. We develop the methodology using a two-step
strategy. First, we let the avatar maker list pairs by intuition,
guessing what possible questions a user may ask to the avatar.
Second, we record actual dialogues between random individuals and
the avatar-maker. We make the Margarita Dialogue Corpus available to
the research community. This corpus comprises the knowledge base in
text format, the video clips for each answer, and the annotated
dialogues.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.60}
}

@InProceedings{dekova:2020:LREC,
  author    = {Dekova, Rositsa},
  title     = {The Ontology of Bulgarian Dialects – Architecture and
Information Retrieval},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4877--4882},
  abstract  = {Following a concise description of the structure, the
paper focuses on the potential of the Ontology of the Bulgarian
Dialects, which demonstrates a novel usage of the ontological
modelling for the purposes of dialect digital archiving and
information processing. The ontology incorporates information on the
dialects of the Bulgarian language and includes data from 84
dialects, spoken not only on the territory of the Republic of
Bulgaria, but also abroad. It encodes both their geographical
distribution and some of their main diagnostic features, such as the
different mutations (also referred to as reflexes) of some of the
```

```
Old Bulgarian vowels. The mutations modelled so far in the ontology
include the reflex of the back nasal vowel /ѫ/ under stress, the
reflex of the back er vowel /ъ/ under stress, and the reflex of the
yat vowel /ѣ/ under stress when it precedes a syllable with a back
vowel. Besides the opportunity for formal structuring of the
considerable amount of data gathered through the years by
dialectologists, the ontology also provides numerous possibilities
for information retrieval – searches by dialect, country, dialect
region, city or village, various combinations of diagnostic
features.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.600}
}

@InProceedings{bonn-EtAl:2020:LREC,
  author    = {Bonn, Julia  and  Palmer, Martha  and  Cai, Zheng
and  Wright-Bettner, Kristin},
  title     = {Spatial AMR: Expanded Spatial Annotation in the
Context of a Grounded Minecraft Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4883--4892},
  abstract  = {This paper presents an expansion to the Abstract
Meaning Representation (AMR) annotation schema that captures fine-
grained semantically and pragmatically derived spatial information
in grounded corpora. We describe a new lexical category
conceptualization and set of spatial annotation tools built in the
context of a multimodal corpus consisting of 170 3D structure-
building dialogues between a human architect and human builder in
Minecraft. Minecraft provides a particularly beneficial spatial
relation-elicitation environment because it automatically tracks
locations and orientations of objects and avatars in the space
according to an absolute Cartesian coordinate system. Through a two-
step process of sentence-level and document-level annotation
designed to capture implicit information, we leverage these
coordinates and bearings in the AMRs in combination with spatial
framework annotation to ground the spatial language in the dialogues
to absolute space.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.601}
}

@InProceedings{klubika-EtAl:2020:LREC,
  author    = {Klubička, Filip  and  Maldonado, Alfredo  and
Mahalunkar, Abhijit  and  Kelleher, John},
  title     = {English WordNet Random Walk Pseudo-Corpora},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {4893--4902},
  abstract  = {This is a resource description paper that describes
the creation and properties of a set of pseudo-corpora generated
artificially from a random walk over the English WordNet taxonomy.
Our WordNet taxonomic random walk implementation allows the
exploration of different random walk hyperparameters and the
generation of a variety of different pseudo-corpora. We find that
different combinations of parameters result in varying statistical
properties of the generated pseudo-corpora. We have published a
total of 81 pseudo-corpora that we have used in our previous
research, but have not exhausted all possible combinations of
hyperparameters, which is why we have also published a codebase that
allows the generation of additional WordNet taxonomic pseudo-corpora
as needed. Ultimately, such pseudo-corpora can be used to train
taxonomic word embeddings, as a way of transferring taxonomic
knowledge into a word embedding space.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.602}
}

@InProceedings{vezzani-dinunzio:2020:LREC,
  author    = {Vezzani, Federica  and  Di Nunzio, Giorgio Maria},
  title     = {On the Formal Standardization of Terminology
Resources: The Case Study of TriMED},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4903--4910},
  abstract  = {The process of standardization plays an important
role in the management of terminological resources. In this context,
we present the work of re-modeling an existing multilingual
terminological database for the medical domain, named TriMED. This
resource was conceived in order to tackle some problems related to
the complexity of medical terminology and to respond to different
users' needs. We provide a methodology that should be followed in
order to make a termbase compliant to the three most recent ISO/TC
37 standards. In particular, we focus on the definition of i) the
structural meta-model of the resource, ii) the data categories
provided, and iii) the TBX format for its implementation. In
addition to the formal standardization of the resource, we describe
the realization of a new data category repository for the management
of the TriMED terminological data and a Web application that can be
used to access the multilingual terminological records.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.603}
}

@InProceedings{alsiyat-piao:2020:LREC,
  author    = {Alsiyat, Israa  and  Piao, Scott},
  title     = {Metaphorical Expressions in Automatic Arabic
Sentiment Analysis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4911--4916},
  abstract  = {Over the recent years, Arabic language resources and
NLP tools have been under rapid development. One of the important
tasks for Arabic natural language processing is the sentiment
analysis. While a significant improvement has been achieved in this
research area, the existing computational models and tools still
suffer from the lack of capability of dealing with Arabic
metaphorical expressions. Metaphor has an important role in Arabic
language due to its unique history and culture. Metaphors provide a
linguistic mechanism for expressing ideas and notions that can be
different from their surface form. Therefore, in order to
efficiently identify true sentiment of Arabic language data, a
computational model needs to be able to “read between lines”. In
this paper, we examine the issue of metaphors in automatic Arabic
sentiment analysis by carrying out an experiment, in which we
observe the performance of a state-of-art Arabic sentiment tool on
metaphors and analyse the result to gain a deeper insight into the
issue. Our experiment evidently shows that metaphors have a
significant impact on the performance of current Arabic sentiment
tools, and it is an important task to develop Arabic language
resources and computational models for Arabic metaphors.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.604}
}

@InProceedings{antognini-faltings:2020:LREC1,
  author    = {Antognini, Diego  and  Faltings, Boi},
  title     = {HotelRec: a Novel Very Large-Scale Hotel
Recommendation Dataset},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4917--4923},
  abstract  = {Today, recommender systems are an inevitable part of
everyone's daily digital routine and are present on most internet
platforms. State-of-the-art deep learning-based models require a
large number of data to achieve their best performance. Many
datasets fulfilling this criterion have been proposed for multiple
domains, such as Amazon products, restaurants, or beers. However,
works and datasets in the hotel domain are limited: the largest
hotel review dataset is below the million samples. Additionally, the
hotel domain suffers from a higher data sparsity than traditional
recommendation datasets and therefore, traditional collaborative-
filtering approaches cannot be applied to such data. In this paper,
we propose HotelRec, a very large-scale hotel recommendation
dataset, based on TripAdvisor, containing 50 million reviews. To the
best of our knowledge, HotelRec is the largest publicly available
dataset in the hotel domain (50M versus 0.9M) and additionally, the
```

largest recommendation dataset in a single domain and with textual reviews (50M versus 22M). We release HotelRec for further research: https://github.com/Diego999/HotelRec.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.605}
}

@InProceedings{vandenberg-EtAl:2020:LREC,
  author   = {van den Berg, Esther and Korfhage, Katharina and Ruppenhofer, Josef and Wiegand, Michael and Markert, Katja},
  title   = {Doctor Who? Framing Through Names and Titles in German},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4924--4932},
  abstract = {Entity framing is the selection of aspects of an entity to promote a particular viewpoint towards that entity. We investigate entity framing of political figures through the use of names and titles in German online discourse, enhancing current research in entity framing through titling and naming that concentrates on English only. We collect tweets that mention prominent German politicians and annotate them for stance. We find that the formality of naming in these tweets correlates positively with their stance. This confirms sociolinguistic observations that naming and titling can have a status-indicating function and suggests that this function is dominant in German tweets mentioning political figures. We also find that this status-indicating function is much weaker in tweets from users that are politically left-leaning than in tweets by right-leaning users. This is in line with observations from moral psychology that left-leaning and right-leaning users assign different importance to maintaining social hierarchies.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.606}
}

@InProceedings{rietzler-EtAl:2020:LREC,
  author   = {Rietzler, Alexander and Stabinger, Sebastian and Opitz, Paul and Engl, Stefan},
  title   = {Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4933--4941},
  abstract = {Aspect-Target Sentiment Classification (ATSC) is a subtask of Aspect-Based Sentiment Analysis (ABSA), which has many applications e.g. in e-commerce, where data and insights from

reviews can be leveraged to create value for businesses and customers. Recently, deep transfer-learning methods have been applied successfully to a myriad of Natural Language Processing (NLP) tasks, including ATSC. Building on top of the prominent BERT language model, we approach ATSC using a two-step procedure: self-supervised domain-specific BERT language model finetuning, followed by supervised task-specific finetuning. Our findings on how to best exploit domain-specific language model finetuning enable us to produce new state-of-the-art performance on the SemEval 2014 Task 4 restaurants dataset. In addition, to explore the real-world robustness of our models, we perform cross-domain evaluation. We show that a cross-domain adapted BERT language model performs significantly better than strong baseline models like vanilla BERT-base and XLNet-base. Finally, we conduct a case study to interpret model prediction errors.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.607}
}

@InProceedings{kang-eshkoltaravella:2020:LREC,
  author   = {Kang, Hyun Jung and Eshkol-Taravella, Iris},
  title   = {An Empirical Examination of Online Restaurant Reviews},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4942--4947},
  abstract = {In the wake of (Pang et al., 2002; Turney, 2002; Liu, 2012) inter alia, opinion mining and sentiment analysis have focused on extracting either positive or negative opinions from texts and determining the targets of these opinions. In this study, we go beyond the coarse-grained positive vs. negative opposition and propose a corpus-based scheme that detects evaluative language at a finer-grained level. We classify each sentence into one of four evaluation types based on the proposed scheme: (1) the reviewer’s opinion on the restaurant (positive, negative, or mixed); (2) the reviewer’s input/feedback to potential customers and restaurant owners (suggestion, advice, or warning) (3) whether the reviewer wants to return to the restaurant (intention); (4) the factual statement about the experience (description). We apply classical machine learning and deep learning methods to show the effectiveness of our scheme. We also interpret the performances that we obtained for each category by taking into account the specificities of the corpus treated.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.608}
}

@InProceedings{kameswari-mamidi:2020:LREC,
  author   = {Kameswari, Lalitha and Mamidi, Radhika},
  title   = {Manovaad: A Novel Approach to Event Oriented Corpus Creation Capturing Subjectivity and Focus},
  booktitle   = {Proceedings of The 12th Language Resources and

Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4948--4954},
  abstract = {In today's era of globalisation, the increased outreach for every event across the world has been leading to conflicting opinions, arguments and disagreements, often reflected in print media and online social platforms. It is necessary to distinguish factual observations from personal judgements in news, as subjectivity in reporting can influence the audience's perception of reality. Several studies conducted on the different styles of reporting in journalism are essential in understanding phenomena such as media bias and multiple interpretations of the same event. This domain finds applications in fields such as Media Studies, Discourse Analysis, Information Extraction, Sentiment Analysis, and Opinion Mining. We present an event corpus Manovaad-v1.0 consisting of 1035 news articles corresponding to 65 events from 3 levels of newspapers viz., Local, National, and International levels. Using this novel format, we correlate the trends in the degree of subjectivity with the geographical closeness of reporting using a Bi-RNN model. We also analyse the role of background and focus in event reporting and capture the focus shift patterns within a global discourse structure for an event. We do this across different levels of reporting and compare the results with the existing work on discourse processing.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.609}
}

@InProceedings{schmidt-minker-werner:2020:LREC,
  author   = {Schmidt, Maria and Minker, Wolfgang and Werner, Steffen},
  title   = {How Users React to Proactive Voice Assistant Behavior While Driving},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {485--490},
  abstract = {Nowadays Personal Assistants (PAs) are available in multiple environments and become increasingly popular to use via voice. Therefore, we aim to provide proactive PA suggestions to car drivers via speech. These suggestions should be neither obtrusive nor increase the drivers’ cognitive load, while enhancing user experience. To assess these factors, we conducted a usability study in which 42 participants perceive proactive voice output in a Wizard-of-Oz study in a driving simulator. Traffic density was varied during a highway drive and it included six in-car-specific use cases. The latter were presented by a proactive voice assistant and in a non-proactive control condition. We assessed the users’ subjective cognitive load and their satisfaction in different

```
questionnaires during the interaction with both PA variants. Furthermore, we analyze the user reactions: both regarding their content and the elapsed response times to PA actions. The results show that proactive assistant behavior is rated similarly positive as non-proactive behavior. Furthermore, the participants agreed to 73.8\% of proactive suggestions. In line with previous research, driving-relevant use cases receive the best ratings, here we reach 82.5\% acceptance. Finally, the users reacted significantly faster to proactive PA actions, which we interpret as less cognitive load compared to non-proactive behavior.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.61}
}

@InProceedings{barhoumi-EtAl:2020:LREC,
  author    = {Barhoumi, Amira  and  Camelin, Nathalie  and  Aloulou, Chafik  and  Estève, Yannick  and  Hadrich Belguith, Lamia},
  title     = {Toward Qualitative Evaluation of Embeddings for Arabic Sentiment Analysis},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4955--4963},
  abstract  = {In this paper, we propose several protocols to evaluate specific embeddings for Arabic sentiment analysis (SA) task. In fact, Arabic language is characterized by its agglutination and morphological richness contributing to great sparsity that could affect embedding quality. This work presents a study that compares embeddings based on words and lemmas in SA frame. We propose first to study the evolution of embedding models trained with different types of corpora (polar and non polar) and explore the variation between embeddings by observing the sentiment stability of neighbors in embedding spaces. Then, we evaluate embeddings with a neural architecture based on convolutional neural network (CNN). We make available our pre-trained embeddings to Arabic NLP research community with free to use. We provide also for free resources used to evaluate our embeddings. Experiments are done on the Large Arabic-Book Reviews (LABR) corpus in binary (positive/negative) classification frame. Our best result reaches 91.9\%, that is higher than the best previous published one (91.5\%).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.610}
}

@InProceedings{morante-EtAl:2020:LREC,
  author    = {Morante, Roser  and  van Son, Chantal  and  Maks, Isa and  Vossen, Piek},
  title     = {Annotating Perspectives on Vaccination},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4964--4973},
  abstract  = {In this paper we present the Vaccination Corpus, a
corpus of texts related to the online vaccination debate that has
been annotated with three layers of information about perspectives:
attribution, claims and opinions. Additionally, events related to
the vaccination debate are also annotated. The corpus contains 294
documents from the Internet which reflect different views on
vaccinations. It has been compiled to study the language of online
debates, with the final goal of experimenting with methodologies to
extract and contrast perspectives in the framework of the
vaccination debate.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.611}
}

@InProceedings{chinearios-francosalvador-benajiba:2020:LREC,
  author    = {Chinea-Rios, Mara  and  Franco-Salvador, Marc  and
Benajiba, Yassine},
  title     = {Aspect On: an Interactive Solution for Post-Editing
the Aspect Extraction based on Online Learning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4974--4981},
  abstract  = {The task of aspect extraction is an important
component of aspect-based sentiment analysis. However, it usually
requires an expensive human post-processing to ensure quality. In
this work we introduce Aspect On, an interactive solution based on
online learning that allows users to post-edit the aspect extraction
with little effort. The Aspect On interface shows the aspects
extracted by a neural model and, given a dataset, annotates its
words with the corresponding aspects. Thanks to the online learning,
Aspect On updates the model automatically and continuously improves
the quality of the aspects displayed to the user. Experimental
results show that Aspect On dramatically reduces the number of user
clicks and effort required to post-edit the aspects extracted by the
model.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.612}
}

@InProceedings{sheoran-EtAl:2020:LREC,
  author    = {Sheoran, Akash  and  Kanojia, Diptesh  and  Joshi,
Aditya  and  Bhattacharyya, Pushpak},
  title     = {Recommendation Chart of Domains for Cross-Domain
Sentiment Analysis: Findings of A 20 Domain Study},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {4982--4990},
  abstract  = {Cross-domain sentiment analysis (CDSA) helps to
address the problem of data scarcity in scenarios where labelled
data for a domain (known as the target domain) is unavailable or
insufficient. However, the decision to choose a domain (known as the
source domain) to leverage from is, at best, intuitive. In this
paper, we investigate text similarity metrics to facilitate source
domain selection for CDSA. We report results on 20 domains (all
possible pairs) using 11 similarity metrics. Specifically, we
compare CDSA performance with these metrics for different domain-
pairs to enable the selection of a suitable source domain, given a
target domain. These metrics include two novel metrics for
evaluating domain adaptability to help source domain selection of
labelled data and utilize word and sentence-based embeddings as
metrics for unlabelled data. The goal of our experiments is a
recommendation chart that gives the K best source domains for CDSA
for a given target domain. We show that the best K source domains
returned by our similarity metrics have a precision of over 50\%,
for varying values of K.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.613}
}

@InProceedings{yan-EtAl:2020:LREC,
  author    = {Yan, Liyun  and  E, Danni  and  Gan, Mei  and
Grouin, Cyril  and  Valette, Mathieu},
  title     = {Inference Annotation of a Chinese Corpus for Opinion
Mining},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {4991--4999},
  abstract  = {Polarity classification (positive, negative or
neutral opinion detection) is well developed in the field of opinion
mining. However, existing tools, which perform with high accuracy on
short sentences and explicit expressions, have limited success
interpreting narrative phrases and inference contexts. In this
article, we will discuss an important aspect of opinion mining:
inference. We will give our definition of inference, classify
different types, provide an annotation framework and analyze the
annotation results. While inferences are often studied in the field
of Natural-language understanding (NLU), we propose to examine
inference as it relates to opinion mining. Firstly, based on
linguistic analysis, we clarify what kind of sentence contains an
inference. We define five types of inference: logical inference,
pragmatic inference, lexical inference, enunciative inference and
discursive inference. Second, we explain our annotation framework
which includes both inference detection and opinion mining. In
short, this manual annotation determines whether or not a target
contains an inference. If so, we then define inference type,
polarity and topic. Using the results of this annotation, we
```

observed several correlation relations which will be used to determine distinctive features for automatic inference classification in further research. We also demonstrate the results of three preliminary classification experiments.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.614}
}

@InProceedings{mohammadi-EtAl:2020:LREC,
  author   = {Mohammadi, Elham and Naji, Nada and Marceau, Louis and Queudot, Marc and Charton, Eric and Kosseim, Leila and Meurs, Marie-Jean},
  title   = {Cooking Up a Neural-based Model for Recipe Classification},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5000--5009},
  abstract = {In this paper, we propose a neural-based model to address the first task of the DEFT 2013 shared task, with the main challenge of a highly imbalanced dataset, using state-of-the-art embedding approaches and deep architectures. We report on our experiments on the use of linguistic features, extracted by Charton et. al. (2014), in different neural models utilizing pretrained embeddings. Our results show that all of the models that use linguistic features outperform their counterpart models that only use pretrained embeddings. The best performing model uses pretrained CamemBERT embeddings as input and CNN as the hidden layer, and uses additional linguistic features. Adding the linguistic features to this model improves its performance by 4.5\% and 11.4\% in terms of micro and macro F1 scores, respectively, leading to state-of-the-art results and an improved classification of the rare classes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.615}
}

@InProceedings{schulder-wiegand-ruppenhofer:2020:LREC,
  author   = {Schulder, Marc and Wiegand, Michael and Ruppenhofer, Josef},
  title   = {Enhancing a Lexicon of Polarity Shifters through the Supervised Classification of Shifting Directions},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5010--5016},
  abstract = {The sentiment polarity of an expression (whether it is perceived as positive, negative or neutral) can be influenced by a number of phenomena, foremost among them negation. Apart from closed-class negation words like "no", "not" or "without", negation can also be caused by so-called polarity shifters. These are content

words, such as verbs, nouns or adjectives, that shift polarities in their opposite direction, e.g. "abandoned" in "abandoned hope" or "alleviate" in "alleviate pain". Many polarity shifters can affect both positive and negative polar expressions, shifting them towards the opposing polarity. However, other shifters are restricted to a single shifting direction. "Recoup" shifts negative to positive in "recoup your losses", but does not affect the positive polarity of "fortune" in "recoup a fortune". Existing polarity shifter lexica only specify whether a word can, in general, cause shifting, but they do not specify when this is limited to one shifting direction. To address this issue we introduce a supervised classifier that determines the shifting direction of shifters. This classifier uses both resource-driven features, such as WordNet relations, and data-driven features like in-context polarity conflicts. Using this classifier we enhance the largest available polarity shifter lexicon.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.616}
}

@InProceedings{regatte-gangula-mamidi:2020:LREC,
  author   = {Regatte, Yashwanth Reddy and Gangula, Rama Rohit Reddy and Mamidi, Radhika},
  title   = {Dataset Creation and Evaluation of Aspect Based Sentiment Analysis in Telugu, a Low Resource Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5017--5024},
  abstract = {In recent years, sentiment analysis has gained popularity as it is essential to moderate and analyse the information across the internet. It has various applications like opinion mining, social media monitoring, and market research. Aspect Based Sentiment Analysis (ABSA) is an area of sentiment analysis which deals with sentiment at a finer level. ABSA classifies sentiment with respect to each aspect to gain greater insights into the sentiment expressed. Significant contributions have been made in ABSA, but this progress is limited only to a few languages with adequate resources. Telugu lags behind in this area of research despite being one of the most spoken languages in India and an enormous amount of data being created each day. In this paper, we create a reliable resource for aspect based sentiment analysis in Telugu. The data is annotated for three tasks namely Aspect Term Extraction, Aspect Polarity Classification and Aspect Categorisation. Further, we develop baselines for the tasks using deep learning methods demonstrating the reliability and usefulness of the resource.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.617}
}

@InProceedings{vrelid-EtAl:2020:LREC,
  author   = {Øvrelid, Lilja and Mæhlum, Petter and Barnes,

```
Jeremy  and  Velldal, Erik},
  title     = {A Fine-grained Sentiment Dataset for Norwegian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5025--5033},
  abstract  = {We here introduce NoReC\_fine, a dataset for fine-
grained sentiment analysis in Norwegian, annotated with respect to
polar expressions, targets and holders of opinion. The underlying
texts are taken from a corpus of professionally authored reviews
from multiple news-sources and across a wide variety of domains,
including literature, games, music, products, movies and more. We
here present a detailed description of this annotation effort. We
provide an overview of the developed annotation guidelines,
illustrated with examples and present an analysis of inter-annotator
agreement. We also report the first experimental results on the
dataset, intended as a preliminary benchmark for further
experiments.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.618}
}

@InProceedings{gong-EtAl:2020:LREC,
  author    = {Gong, Xiaochang  and  Zhao, Qin  and  Zhang, Jun  and
Mao, Ruibin  and  Xu, Ruifeng},
  title     = {The Design and Construction of a Chinese Sarcasm
Dataset},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5034--5039},
  abstract  = {As a typical multi-layered semi-conscious language
phenomenon, sarcasm is widely existed in social media text for
enhancing the emotion expression. Thus, the detection and processing
of sarcasm is important to social media analysis.However, most
existing sarcasm dataset are in English and there is still a lack of
authoritative Chinese sarcasm dataset. In this paper, we presents
the design and construction of a largest high-quality Chinese
sarcasm dataset, which contains 2,486 manual annotated sarcastic
texts and 89,296 non-sarcastic texts. Furthermore, a balanced
dataset through elaborately sampling the same amount non-sarcastic
texts for training sarcasm classifier. Using the dataset as the
benchmark, some sarcasm classification methods are evaluated.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.619}
}

@InProceedings{asai-EtAl:2020:LREC,
  author    = {Asai, Sara  and  Yoshino, Koichiro  and  Shinagawa,
Seitaro  and  Sakti, Sakriani  and  Nakamura, Satoshi},
```

```
  title     = {Emotional Speech Corpus for Persuasive Dialogue
System},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {491--497},
  abstract  = {Expressing emotion is known as an efficient way to
persuade one’s dialogue partner to accept one’s claim or proposal.
Emotional expression in speech can express the speaker’s emotion
more directly than using only emotion expression in the text, which
will lead to a more persuasive dialogue. In this paper, we built a
speech dialogue corpus in a persuasive scenario that uses emotional
expressions to build a persuasive dialogue system with emotional
expressions. We extended an existing text dialogue corpus by adding
variations of emotional responses to cover different combinations of
broad dialogue context and a variety of emotional states by crowd-
sourcing. Then, we recorded emotional speech consisting of of
collected emotional expressions spoken by a voice actor. The
experimental results indicate that the collected emotional
expressions with their speeches have higher emotional expressiveness
for expressing the system’s emotion to users.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.62}
}

@InProceedings{yuan-EtAl:2020:LREC,
  author    = {Yuan, Chaofa  and  Liu, Yuhan  and  Yin, Rongdi  and
Zhang, Jun  and  Zhu, Qinling  and  Mao, Ruibin  and  Xu, Ruifeng},
  title     = {Target-based Sentiment Annotation in Chinese
Financial News},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5040--5045},
  abstract  = {This paper presents the design and construction of a
large-scale target-based sentiment annotation corpus on Chinese
financial news text. Different from the most existing paragraph/
document-based annotation corpus, in this study, target-based fine-
grained sentiment annotation is performed. The companies, brands and
other financial entities are regarded as the targets. The clause
reflecting the profitability, loss or other business status of
financial entities is regarded as the sentiment expression for
determining the polarity. Based on high quality annotation guideline
and effective quality control strategy, a corpus with 8,314 target-
level sentiment annotation is constructed on 6,336 paragraphs from
Chinese financial news text. Based on this corpus, several state-of-
the-art sentiment analysis models are evaluated.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.620}
}
```

```
@InProceedings{-EtAl:2020:LREC,
  author    = {., Mamta  and  Ekbal, Asif  and  Bhattacharyya, Pushpak  and  Srivastava, Shikha  and  Kumar, Alka  and  Saha, Tista},
  title     = {Multi-domain Tweet Corpora for Sentiment Analysis: Resource Creation and Evaluation},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5046--5054},
  abstract  = {Due to the phenomenal growth of online content in recent time, sentiment analysis has attracted attention of the researchers and developers. A number of benchmark annotated corpora are available for domains like movie reviews, product reviews, hotel reviews, etc.The pervasiveness of social media has also lead to a huge amount of content posted by users who are misusing the power of social media to spread false beliefs and to negatively influence others. This type of content is coming from the domains like terrorism, cybersecurity, technology, social issues, etc. Mining of opinions from these domains is important to create a socially intelligent system to provide security to the public and to maintain the law and order situations. To the best of our knowledge, there is no publicly available tweet corpora for such pervasive domains. Hence, we firstly create a multi-domain tweet sentiment corpora and then establish a deep neural network based baseline framework to address the above mentioned issues. Annotated corpus has Cohen's Kappa measurement for annotation quality of 0.770, which shows that the data is of acceptable quality. We are able to achieve 84.65\% accuracy for sentiment analysis by using an ensemble of Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit(GRU).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.621}
}

@InProceedings{antniorodrigues-EtAl:2020:LREC,
  author    = {António Rodrigues, João  and  Branco, Ruben  and  Silva, João  and  Branco, António},
  title     = {Reproduction and Revival of the Argument Reasoning Comprehension Task},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5055--5064},
  abstract  = {Reproduction of scientific findings is essential for scientific development across all scientific disciplines and reproducing results of previous works is a basic requirement for validating the hypothesis and conclusions put forward by them. This
```

paper reports on the scientific reproduction of several systems addressing the Argument Reasoning Comprehension Task of SemEval2018. Given a recent publication that pointed out spurious statistical cues in the data set used in the shared task, and that produced a revised version of it, we also evaluated the reproduced systems with this new data set. The exercise reported here shows that, in general, the reproduction of these systems is successful with scores in line with those reported in SemEval2018. However, the performance scores are worst than those, and even below the random baseline, when the reproduced systems are run over the revised data set expunged from data artifacts. This demonstrates that this task is actually a much harder challenge than what could have been perceived from the inflated, close to human-level performance scores obtained with the data set used in SemEval2018. This calls for a revival of this task as there is much room for improvement until systems may come close to the upper bound provided by human performance.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.622}
}

@InProceedings{morenoortiz-fernandezcruz-hernndez:2020:LREC,
  author   = {Moreno-Ortiz, Antonio and Fernandez-Cruz, Javier and Hernández, Chantal Pérez Chantal},
  title   = {Design and Evaluation of SentiEcon: a fine-grained Economic/Financial Sentiment Lexicon from a Corpus of Business News},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5065--5072},
  abstract = {In this paper we present, describe, and evaluate SentiEcon, a large, comprehensive, domain-specific computational lexicon designed for sentiment analysis applications, for which we compiled our own corpus of online business news. SentiEcon was created as a plug-in lexicon for the sentiment analysis tool Lingmotif, and thus it follows its data structure requirements and presupposes the availability of a general-language core sentiment lexicon that covers non-specific sentiment-carrying terms and phrases. It contains 6,470 entries, both single and multi-word expressions, each with tags denoting their semantic orientation and intensity. We evaluate SentiEcon’s performance by comparing results in a sentence classification task using exclusively sentiment words as features. This sentence dataset was extracted from business news texts, and included certain key words known to recurrently convey strong semantic orientation, such as “debt”, “inflation” or “markets”. The results show that performance is significantly improved when adding SentiEcon to the general-language sentiment lexicon.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.623}
}

@InProceedings{abercrombie-batistanavarro:2020:LREC,

```
  author    = {Abercrombie, Gavin  and  Batista-Navarro, Riza},
  title     = {ParlVote: A Corpus for Sentiment Analysis of
Political Debates},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5073--5078},
  abstract  = {Debate transcripts from the UK Parliament contain
information about the positions taken by politicians towards
important topics, but are difficult for people to process manually.
While sentiment analysis of debate speeches could facilitate
understanding of the speakers' stated opinions, datasets currently
available for this task are small when compared to the benchmark
corpora in other domains. We present ParlVote, a new, larger corpus
of parliamentary debate speeches for use in the evaluation of
sentiment analysis systems for the political domain. We also perform
a number of initial experiments on this dataset, testing a variety
of approaches to the classification of sentiment polarity in debate
speeches. These include a linear classifier as well as a neural
network trained using a transformer word embedding model (BERT), and
fine-tuned on the parliamentary speeches. We find that in many
scenarios, a linear classifier trained on a bag-of-words text
representation achieves the best results. However, with the largest
dataset, the transformer-based model combined with a neural
classifier provides the best performance. We suggest that further
experimentation with classification models and observations of the
debate content and structure are required, and that there remains
much room for improvement in parliamentary sentiment analysis.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.624}
}

@InProceedings{tian-kbler:2020:LREC,
  author    = {Tian, Zuoyu  and  Kübler, Sandra},
  title     = {Offensive Language Detection Using Brown Clustering},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5079--5087},
  abstract  = {In this study, we investigate the use of Brown
clustering for offensive language detection. Brown clustering has
been shown to be of little use when the task involves distinguishing
word polarity in sentiment analysis tasks. In contrast to previous
work, we train Brown clusters separately on positive and negative
sentiment data, but then combine the information into a single
complex feature per word. This way of representing words results in
stable improvements in offensive language detection, when used as
the only features or in combination with words or character n-grams.
Brown clusters add important information, even when combined with
```

words or character n-grams or with standard word embeddings in a convolutional neural network. However, we also found different trends between the two offensive language data sets we used.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.625}
}

@InProceedings{assimakopoulos-EtAl:2020:LREC,
  author   = {Assimakopoulos, Stavros and Vella Muskat, Rebecca and van der Plas, Lonneke and Gatt, Albert},
  title   = {Annotating for Hate Speech: The MaNeCo Corpus and Some Input from Critical Discourse Analysis},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5088--5097},
  abstract = {This paper presents a novel scheme for the annotation of hate speech in corpora of Web 2.0 commentary. The proposed scheme is motivated by the critical analysis of posts made in reaction to news reports on the Mediterranean migration crisis and LGBTIQ+ matters in Malta, which was conducted under the auspices of the EU-funded C.O.N.T.A.C.T. project. Based on the realisation that hate speech is not a clear-cut category to begin with, appears to belong to a continuum of discriminatory discourse and is often realised through the use of indirect linguistic means, it is argued that annotation schemes for its detection should refrain from directly including the label 'hate speech,' as different annotators might have different thresholds as to what constitutes hate speech and what not. In view of this, we propose a multi-layer annotation scheme, which is pilot-tested against a binary ±hate speech classification and appears to yield higher inter-annotator agreement. Motivating the postulation of our scheme, we then present the MaNeCo corpus on which it will eventually be used; a substantial corpus of on-line newspaper comments spanning 10 years.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.626}
}

@InProceedings{cignarella-EtAl:2020:LREC,
  author   = {Cignarella, Alessandra Teresa and Sanguinetti, Manuela and Bosco, Cristina and Rosso, Paolo},
  title   = {Marking Irony Activators in a Universal Dependencies Treebank: The Case of an Italian Twitter Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5098--5105},
  abstract = {The recognition of irony is a challenging task in the domain of Sentiment Analysis, and the availability of annotated corpora may be crucial for its automatic processing. In this paper

```
we describe a fine-grained annotation scheme centered on irony, in
which we highlight the tokens that are responsible for its
activation, (irony activators) and their morpho-syntactic features.
As our case study we therefore introduce a recently released
Universal Dependencies treebank for Italian which includes ironic
tweets: TWITTIRÒ-UD. For the purposes of this study, we enriched the
existing annotation in the treebank, with a further level that
includes irony activators. A description and discussion of the
annotation scheme is provided with a definition of irony activators
and the guidelines for their annotation. This qualitative study on
the different layers of annotation applied on the same dataset can
shed some light on the process of human annotation, and irony
annotation in particular, and on the usefulness of this
representation for developing computational models of irony to be
used for training purposes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.627}
}

@InProceedings{chiruzzo-castro-ros:2020:LREC,
  author    = {Chiruzzo, Luis  and  Castro, Santiago  and  Rosá,
Aiala},
  title     = {HAHA 2019 Dataset: A Corpus for Humor Analysis in
Spanish},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5106--5112},
  abstract  = {This paper presents the development of a corpus of
30,000 Spanish tweets that were crowd-annotated with humor value and
funniness score. The corpus contains approximately 38.6\% of
humorous tweets with an average score of 2.04 in a scale from 1 to 5
for the humorous tweets. The corpus has been used in an automatic
humor recognition and analysis competition, obtaining encouraging
results from the participants.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.628}
}

@InProceedings{pitenis-zampieri-ranasinghe:2020:LREC,
  author    = {Pitenis, Zesis  and  Zampieri, Marcos  and
Ranasinghe, Tharindu},
  title     = {Offensive Language Identification in Greek},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5113--5119},
  abstract  = {As offensive language has become a rising issue for
online communities and social media platforms, researchers have been
investigating ways of coping with abusive content and developing
```

```
systems to detect its different types: cyberbullying, hate speech,
aggression, etc. With a few notable exceptions, most research on
this topic so far has dealt with English. This is mostly due to the
availability of language resources for English. To address this
shortcoming, this paper presents the first Greek annotated dataset
for offensive language identification: the Offensive Greek Tweet
Dataset (OGTD). OGTD is a manually annotated dataset containing
4,779 posts from Twitter annotated as offensive and not offensive.
Along with a detailed description of the dataset, we evaluate
several computational models trained and tested on this data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.629}
}

@InProceedings{bangalorekantharaju-EtAl:2020:LREC,
  author    = {Bangalore Kantharaju, Reshmashree  and  Langlet,
Caroline  and  Barange, Mukesh  and  Clavel, Chloé  and  Pelachaud,
Catherine},
  title     = {Multimodal Analysis of Cohesion in Multi-party
Interactions},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {498--507},
  abstract  = {Group cohesion is an emergent phenomenon that
describes the tendency of the group members' shared commitment to
group tasks and the interpersonal attraction among them. This paper
presents a multimodal analysis of group cohesion using a corpus of
multi-party interactions. We utilize 16 two-minute segments
annotated with cohesion from the AMI corpus. We define three layers
of modalities: non-verbal social cues, dialogue acts and
interruptions. The initial analysis is performed at the individual
level and later, we combine the different modalities to observe
their impact on perceived level of cohesion. Results indicate that
occurrence of laughter and interruption are higher in high cohesive
segments. We also observe that, dialogue acts and head nods did not
have an impact on the level of cohesion by itself. However, when
combined there was an impact on the perceived level of cohesion.
Overall, the analysis shows that multimodal cues are crucial for
accurate analysis of group cohesion.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.63}
}

@InProceedings{bick:2020:LREC1,
  author    = {Bick, Eckhard},
  title     = {Syntax and Semantics in a Treebank for Esperanto},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {5120--5127},
  abstract  = {In this paper we describe and evaluate syntactic and
semantic aspects of Arbobanko, a treebank for the artificial
language Esperanto, as well as tools and methods used in the
production of the treebank. In addition to classical morphosyntax
and dependency structure, the treebank was enriched with a lexical-
semantic layer covering named entities, a semantic type ontology for
nouns and adjectives and a framenet-inspired semantic classification
of verbs. For an under-resourced language, the quality of automatic
syntactic and semantic pre-annotation is of obvious importance, and
by evaluating the underlying parser and the coverage of its semantic
ontologies, we try to answer the question whether the language's
extremely regular morphology and transparent semantic affixes
translate into a more regular syntax and higher parsing accuracy. On
the linguistic side, the treebank allows us to address and quantify
typological issues such as the question of word order, auxiliary
constructions, lexical transparency and semantic type ambiguity in
Esperanto.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.630}
}

@InProceedings{dione:2020:LREC,
  author    = {Dione, Cheikh M. Bamba},
  title     = {Implementation and Evaluation of an LFG-based Parser
for Wolof},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5128--5136},
  abstract  = {This paper reports on a parsing system for Wolof
based on the LFG formalism. The parser covers core constructions of
Wolof, including noun classes, cleft, copula, causative and
applicative sentences. It also deals with several types of
coordination, including same constituent coordination, asymmetric
and asyndetic coordination. The system uses a cascade of finite-
state transducers for word tokenization and morphological analysis
as well as various lexicons. In addition, robust parsing techniques,
including fragmenting and skimming, are used to optimize grammar
coverage. Parsing coverage is evaluated by running test-suites of
naturally occurring Wolof sentences through the parser. The
evaluation of parsing coverage reveals that 72.72\% of the test
sentences receive full parses; 27.27\% receive partial parses. To
measure accuracy, the parsed sentences are disambiguated manually
using an incremental parsebanking approach based on discriminants.
The evaluation of parsing quality reveals that the parser achieves
67.2\% recall, 92.8\% precision and an f-score of 77.9\%.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.631}
}

@InProceedings{hellwig-EtAl:2020:LREC,
  author    = {Hellwig, Oliver  and  Scarlata, Salvatore  and
```

```
Ackermann, Elia  and  Widmer, Paul},
  title     = {The Treebank of Vedic Sanskrit},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5137--5146},
  abstract  = {This paper introduces the first treebank of Vedic
Sanskrit, a morphologically rich ancient Indian language that is of
central importance for linguistic and historical research. The
selection of the more than 3,700 sentences contained in this
treebank reflects the development of metrical and prose texts over a
period of 600 years. We discuss how these sentences are annotated in
the Universal Dependencies scheme and which syntactic constructions
required special attention. In addition, we describe a syntactic
labeler based on neural networks that supports the initial
annotation of the treebank, and whose evaluation can be helpful for
setting up a full syntactic parser of Vedic Sanskrit.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.632}
}

@InProceedings{anderson-gmezrodrguez:2020:LREC,
  author    = {Anderson, Mark  and  Gómez-Rodríguez, Carlos},
  title     = {Inherent Dependency Displacement Bias of Transition-
Based Algorithms},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5147--5155},
  abstract  = {A wide variety of transition-based algorithms are
currently used for dependency parsers. Empirical studies have shown
that performance varies across different treebanks in such a way
that one algorithm outperforms another on one treebank and the
reverse is true for a different treebank. There is often no
discernible reason for what causes one algorithm to be more suitable
for a certain treebank and less so for another. In this paper we
shed some light on this by introducing the concept of an algorithm's
inherent dependency displacement distribution. This characterises
the bias of the algorithm in terms of dependency displacement, which
quantify both distance and direction of syntactic relations. We show
that the similarity of an algorithm's inherent distribution to a
treebank's displacement distribution is clearly correlated to the
algorithm's parsing performance on that treebank, specificially with
highly significant and substantial correlations for the predominant
sentence lengths in Universal Dependency treebanks. We also obtain
results which show a more discrete analysis of dependency
displacement does not result in any meaningful correlations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.633}
}
```

```
@InProceedings{kayadelen-ozturel-bohnet:2020:LREC,
  author    = {Kayadelen, Tolga  and  Ozturel, Adnan  and  Bohnet,
Bernd},
  title     = {A Gold Standard Dependency Treebank for Turkish},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5156--5163},
  abstract  = {We introduce TWT; a new treebank for Turkish which
consists of web and Wikipedia sentences that are annotated for
segmentation, morphology, part-of-speech and dependency relations.
To date, it is the largest publicly available human-annotated
morpho-syntactic Turkish treebank in terms of the annotated word
count. It is also the first large Turkish dependency treebank that
has a dedicated Wikipedia section. We present the tagsets and the
methodology that are used in annotating the treebank and also the
results of the baseline experiments on Turkish dependency parsing
with this treebank.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.634}
}

@InProceedings{eshkoltaravella-EtAl:2020:LREC,
  author    = {Eshkol-Taravella, Iris  and  Maarouf, Mariame  and
Badin, Flora  and  Skrovec, Marie  and  Tellier, Isabelle},
  title     = {Chunk Different Kind of Spoken Discourse: Challenges
for Machine Learning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5164--5168},
  abstract  = {This paper describes the development of a chunker for
spoken data by supervised machine learning using the CRFs, based on
a small reference corpus composed of two kinds of discourse:
prepared monologue vs. spontaneous talk in interaction. The
methodology considers the specific character of the spoken data. The
machine learning uses the results of several available taggers,
without correcting the results manually. Experiments show that the
discourse type (monologue vs. free talk), the speech nature
(spontaneous vs. prepared) and the corpus size can influence the
results of the machine learning process and must be considered while
interpreting the results.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.635}
}

@InProceedings{falenska-EtAl:2020:LREC,
  author    = {Falenska, Agnieszka  and  Czesznak, Zoltán  and
Jung, Kerstin  and  Völkel, Moritz  and  Seeker, Wolfgang  and
```

```
Kuhn, Jonas},
  title     = {GRAIN-S: Manually Annotated Syntax for German
Interviews},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5169--5177},
  abstract  = {We present GRAIN-S, a set of manually created
syntactic annotations for radio interviews in German. The dataset
extends an existing corpus GRAIN and comes with constituency and
dependency trees for six interviews. The rare combination of gold-
and silver-standard annotation layers coming from GRAIN with high-
quality syntax trees can serve as a useful resource for speech- and
text-based research. Moreover, since interviews can be put between
carefully prepared speech and spontaneous conversational speech,
they cover phenomena not seen in traditional newspaper-based
treebanks. Therefore, GRAIN-S can contribute to research into
techniques for model adaptation and for building more corpus-
independent tools. GRAIN-S follows TIGER, one of the established
syntactic treebanks of German. We describe the annotation process
and discuss decisions necessary to adapt the original TIGER
guidelines to the interviews domain. Next, we give details on the
conversion from TIGER-style trees to dependency trees. We provide
data statistics and demonstrate differences between the new dataset
and existing out-of-domain test sets annotated with TIGER syntactic
structures. Finally, we provide baseline parsing results for further
comparison.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.636}
}

@InProceedings{ishola-zeman:2020:LREC,
  author    = {Ishola, Olájídé  and  Zeman, Daniel},
  title     = {Yorùbá Dependency Treebank (YTB)},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5178--5186},
  abstract  = {Low-resource languages present enormous NLP
opportunities as well as varying degrees of difficulties. The newly
released treebank of hand-annotated parts of the Yoruba Bible
provides an avenue for dependency analysis of the Yoruba language;
the application of a new grammar formalism to the language. In this
paper, we discuss our choice of Universal Dependencies, important
dependency annotation decisions considered in the creation of the
first annotation guidelines for Yoruba and results of our parsing
experiments. We also lay the foundation for future incorporation of
other domains with the initial test on Yoruba Wikipedia articles and
highlighted future directions for the rapid expansion of the
```

treebank.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.637}
}

@InProceedings{yamakata-mori-carroll:2020:LREC,
  author   = {Yamakata, Yoko and Mori, Shinsuke and Carroll, John},
  title   = {English Recipe Flow Graph Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5187--5194},
  abstract = {We present an annotated corpus of English cooking recipe procedures, and describe and evaluate computational methods for learning these annotations. The corpus consists of 300 recipes written by members of the public, which we have annotated with domain-specific linguistic and semantic structure. Each recipe is annotated with (1) `recipe named entities' (r-NEs) specific to the recipe domain, and (2) a flow graph representing in detail the sequencing of steps, and interactions between cooking tools, food ingredients and the products of intermediate steps. For these two kinds of annotations, inter-annotator agreement ranges from 82.3 to 90.5 F1, indicating that our annotation scheme is appropriate and consistent. We experiment with producing these annotations automatically. For r-NE tagging we train a deep neural network NER tool; to compute flow graphs we train a dependency-style parsing procedure which we apply to the entire sequence of r-NEs in a recipe.In evaluations, our systems achieve 71.1 to 87.5 F1, demonstrating that our annotation scheme is learnable.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.638}
}

@InProceedings{kubota-EtAl:2020:LREC,
  author   = {Kubota, Yusuke and Mineshima, Koji and Hayashi, Noritsugu and Okano, Shinya},
  title   = {Development of a General-Purpose Categorial Grammar Treebank},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5195--5201},
  abstract = {This paper introduces ABC Treebank, a general-purpose categorial grammar (CG) treebank for Japanese. It is 'general-purpose' in the sense that it is not tailored to a specific variant of CG, but rather aims to offer a theory-neutral linguistic resource (as much as possible) which can be converted to different versions of CG (specifically, CCG and Type-Logical Grammar) relatively easily. In terms of linguistic analysis, it improves over the

existing Japanese CG treebank (Japanese CCGBank) on the treatment of certain linguistic phenomena (passives, causatives, and control/ raising predicates) for which the lexical specification of the syntactic information reflecting local dependencies turns out to be crucial. In this paper, we describe the underlying 'theory' dubbed ABC Grammar that is taken as a basis for our treebank, outline the general construction of the corpus, and report on some preliminary results applying the treebank in a semantic parsing system for generating logical representations of sentences.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.639}
}

@InProceedings{nedelchev-usbeck-lehmann:2020:LREC,
  author   = {Nedelchev, Rostislav and Usbeck, Ricardo and Lehmann, Jens},
  title   = {Treating Dialogue Quality Evaluation as an Anomaly Detection Problem},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {508--512},
  abstract = {Dialogue systems for interaction with humans have been enjoying increased popularity in the research and industry fields. To this day, the best way to estimate their success is through means of human evaluation and not automated approaches, despite the abundance of work done in the field. In this paper, we investigate the effectiveness of perceiving dialogue evaluation as an anomaly detection task. The paper looks into four dialogue modeling approaches and how their objective functions correlate with human annotation scores. A high-level perspective exhibits negative results. However, a more in-depth look shows some potential for using anomaly detection for evaluating dialogues.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.64}
}

@InProceedings{ehsan-butt:2020:LREC,
  author   = {Ehsan, Toqeer and Butt, Miriam},
  title   = {Dependency Parsing for Urdu: Resources, Conversions and Learning},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5202--5207},
  abstract = {This paper adds to the available resources for the under-resourced language Urdu by converting different types of existing treebanks for Urdu into a common format that is based on Universal Dependencies. We present comparative results for training two dependency parsers, the MaltParser and a transition-based BiLSTM

```
parser on this new resource. The BiLSTM parser incorporates word
embeddings which improve the parsing results significantly. The
BiLSTM parser outperforms the MaltParser with a UAS of 89.6 and an
LAS of 84.2 with respect to our standardized treebank resource.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.640}
}

@InProceedings{hajic-EtAl:2020:LREC,
  author    = {Hajic, Jan  and  Bejček, Eduard  and  Hlavacova,
Jaroslava  and  Mikulová, Marie  and  Straka, Milan  and  Štěpánek,
Jan  and  Štěpánková, Barbora},
  title     = {Prague Dependency Treebank - Consolidated 1.0},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5208--5218},
  abstract  = {We present a richly annotated and genre-diversified
language resource, the Prague Dependency Treebank-Consolidated 1.0
(PDT-C 1.0), the purpose of which is - as it always been the case
for the family of the Prague Dependency Treebanks - to serve both as
a training data for various types of NLP tasks as well as for
linguistically-oriented research. PDT-C 1.0 contains four different
datasets of Czech, uniformly annotated using the standard PDT scheme
(albeit not everything is annotated manually, as we describe in
detail here). The texts come from different sources: daily newspaper
articles, Czech translation of the Wall Street Journal, transcribed
dialogs and a small amount of user-generated, short, often non-
standard language segments typed into a web translator. Altogether,
the treebank contains around 180,000 sentences with their
morphological, surface and deep syntactic annotation. The diversity
of the texts and annotations should serve well the NLP applications
as well as it is an invaluable resource for linguistic research,
including comparative studies regarding texts of different genres.
The corpus is publicly and freely available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.641}
}

@InProceedings{johansson-adesam:2020:LREC,
  author    = {Johansson, Richard  and  Adesam, Yvonne},
  title     = {Training a Swedish Constituency Parser on Six
Incompatible Treebanks},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5219--5224},
  abstract  = {We investigate a transition-based parser that uses
Eukalyptus, a function-tagged constituent treebank for Swedish which
includes discontinuous constituents. In addition, we show that the
```

```
accuracy of this parser can be improved by using a multitask
learning architecture that makes it possible to train the parser on
additional treebanks that use other annotation models.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.642}
}

@InProceedings{vacareanu-EtAl:2020:LREC,
  author    = {Vacareanu, Robert  and  Gouveia Barbosa, George
Caique  and  Valenzuela-Escárcega, Marco A.  and  Surdeanu, Mihai},
  title     = {Parsing as Tagging},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5225--5231},
  abstract  = {We propose a simple yet accurate method for
dependency parsing that treats parsing as tagging (PaT). That is,
our approach addresses the parsing of dependency trees with a
sequence model implemented with a bidirectional LSTM over BERT
embeddings, where the “tag” to be predicted at each token position
is the relative position of the corresponding head. For example, for
the sentence John eats cake, the tag to be predicted for the token
cake is -1 because its head (eats) occurs one token to the left.
Despite its simplicity, our approach performs well. For example, our
approach outperforms the state-of-the-art method of (Fernández-
González and Gómez-Rodríguez, 2019) on Universal Dependencies (UD)
by 1.76\% unlabeled attachment score (UAS) for English, 1.98\% UAS
for French, and 1.16\% UAS for German. On average, on 12 UD
languages, our method with minimal tuning performs comparably with
this state-of-the-art approach: better by 0.11\% UAS, and worse by
0.58\% LAS.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.643}
}

@InProceedings{bouma-EtAl:2020:LREC,
  author    = {Bouma, Gerlof  and  Coussé, Evie  and  Dijkstra,
Trude  and  van der Sijs, Nicoline},
  title     = {The EDGeS Diachronic Bible Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5232--5239},
  abstract  = {We present the EDGeS Diachronic Bible Corpus: a
diachronically and synchronically parallel corpus of Bible
translations in Dutch, English, German and Swedish, with texts from
the 14th century until today. It is compiled in the context of an
intended longitudinal and contrastive study of complex verb
constructions in Germanic. The paper discusses the corpus design
principles, its selection of 36 Bibles, and the information and
```

metadata encoded for the corpus texts. The EDGeS corpus will be available in two forms: the whole corpus will be accessible for researchers behind a login in the well-known OPUS search infrastructure, and the open subpart of the corpus will be available for download.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.644}
}

@InProceedings{sanguinetti-EtAl:2020:LREC,
  author   = {Sanguinetti, Manuela and Bosco, Cristina and Cassidy, Lauren and Çetinoğlu, Özlem and Cignarella, Alessandra Teresa and Lynn, Teresa and Rehbein, Ines and Ruppenhofer, Josef and Seddah, Djamé and Zeldes, Amir},
  title   = {Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5240--5250},
  abstract = {The paper presents a discussion on the main linguistic phenomena of user-generated texts found in web and social media, and proposes a set of annotation guidelines for their treatment within the Universal Dependencies (UD) framework. Given on the one hand the increasing number of treebanks featuring user-generated content, and its somewhat inconsistent treatment in these resources on the other, the aim of this paper is twofold: (1) to provide a short, though comprehensive, overview of such treebanks - based on available literature - along with their main features and a comparative analysis of their annotation criteria, and (2) to propose a set of tentative UD-based annotation guidelines, to promote consistent treatment of the particular phenomena found in these types of texts. The main goal of this paper is to provide a common framework for those teams interested in developing similar resources in UD, thus enabling cross-linguistic consistency, which is a principle that has always been in the spirit of UD.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.645}
}

@InProceedings{berdicevskis-eckhoff:2020:LREC,
  author   = {Berdicevskis, Aleksandrs and Eckhoff, Hanne},
  title   = {A Diachronic Treebank of Russian Spanning More Than a Thousand Years},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5251--5256},
  abstract = {We describe the Tromsø Old Russian and Old Church Slavonic Treebank (TOROT) that spans from the earliest Old Church

```
Slavonic to modern Russian texts, covering more than a thousand
years of continuous language history. We focus on the latest
additions to the treebank, first of all, the modern subcorpus that
was created by a high-quality conversion of the existing treebank of
contemporary standard Russian (SynTagRus).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.646}
}

@InProceedings{kogkalidis-moortgat-moot:2020:LREC,
  author    = {Kogkalidis, Konstantinos  and  Moortgat, Michael  and
Moot, Richard},
  title     = {ÆTHEL: Automatically Extracted Typelogical
Derivations for Dutch},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5257--5266},
  abstract  = {We present ÆTHEL, a semantic compositionality dataset
for written Dutch. ÆTHEL consists of two parts. First, it contains a
lexicon of supertags for about 900 000 words in context. The
supertags correspond to types of the simply typed linear lambda-
calculus, enhanced with dependency decorations that capture
grammatical roles supplementary to function-argument structures. On
the basis of these types, ÆTHEL further provides 72 192 validated
derivations, presented in four formats: natural-deduction and
sequent-style proofs, linear logic proofnets and the associated
programs (lambda terms) for meaning composition. ÆTHEL's types and
derivations are obtained by means of an extraction algorithm applied
to the syntactic analyses of LASSY Small, the gold standard corpus
of written Dutch. We discuss the extraction algorithm and show how
`virtual elements' in the original LASSY annotation of unbounded
dependencies and coordination phenomena give rise to higher-order
types. We suggest some example usecases highlighting the benefits of
a type-driven approach at the syntax semantics interface. The
following resources are open-sourced with ÆTHEL: the lexical
mappings between words and types, a subset of the dataset consisting
of 7 924 semantic parses, and the Python code that implements the
extraction algorithm.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.647}
}

@InProceedings{gessler-EtAl:2020:LREC,
  author    = {Gessler, Luke  and  Peng, Siyao  and  Liu, Yang  and
Zhu, Yilun  and  Behzad, Shabnam  and  Zeldes, Amir},
  title     = {GUMBY – A Free, Balanced, and Rich English Web
Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {5267--5275},
  abstract  = {We present a freely available, genre-balanced English
web corpus totaling 4M tokens and featuring a large number of high-
quality automatic annotation layers, including dependency trees,
non-named entity annotations, coreference resolution, and discourse
trees in Rhetorical Structure Theory. By tapping open online data
sources the corpus is meant to offer a more sizable alternative to
smaller manually created annotated data sets, while avoiding
pitfalls such as imbalanced or unknown composition, licensing
problems, and low-quality natural language processing. We harness
knowledge from multiple annotation layers in order to achieve a
"better than NLP" benchmark and evaluate the accuracy of the
resulting resource.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.648}
}

@InProceedings{quasthoff-EtAl:2020:LREC,
  author    = {Quasthoff, Uwe  and  Hellan, Lars  and  Körner, Erik
and  Eckart, Thomas  and  Goldhahn, Dirk  and  Beermann, Dorothee},
  title     = {Typical Sentences as a Resource for Valence},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5276--5281},
  abstract  = {Verb valence information can be derived from corpora
by using subcorpora of typical sentences that are constructed in a
language independent manner based on frequent POS structures. The
inspection of typical sentences with a fixed verb in a certain
position can show the valence information directly. Using verb
fingerprints, consisting of the most typical sentence patterns the
verb appears in, we are able to identify standard valence patterns
and compare them against a language's valence profile. With a very
limited number of training data per language, valence information
for other verbs can be derived as well. Based on the Norwegian
valence patterns we are able to find comparative patterns in German
where typical sentences are able to express the same situation in an
equivalent way and can so construct verb valence pairs for a
bilingual PolyVal dictionary. This contribution discusses this
application with a focus on the Norwegian valence dictionary
NorVal.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.649}
}

@InProceedings{rach-EtAl:2020:LREC,
  author    = {Rach, Niklas  and  Matsuda, Yuki  and  Daxenberger,
Johannes  and  Ultes, Stefan  and  Yasumoto, Keiichi  and  Minker,
Wolfgang},
  title     = {Evaluation of Argument Search Approaches in the
Context of Argumentative Dialogue Systems},
  booktitle      = {Proceedings of The 12th Language Resources and
```

Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {513--522},
  abstract = {We present an approach to evaluate argument search techniques in view of their use in argumentative dialogue systems by assessing quality aspects of the retrieved arguments. To this end, we introduce a dialogue system that presents arguments by means of a virtual avatar and synthetic speech to users and allows them to rate the presented content in four different categories (Interesting, Convincing, Comprehensible, Relation). The approach is applied in a user study in order to compare two state of the art argument search engines to each other and with a system based on traditional web search. The results show a significant advantage of the two search engines over the baseline. Moreover, the two search engines show significant advantages over each other in different categories, thereby reflecting strengths and weaknesses of the different underlying techniques.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.65}
}

@InProceedings{hildebrand-hemati-mehler:2020:LREC,
  author   = {Hildebrand, Jonathan and Hemati, Wahed and Mehler, Alexander},
  title   = {Recognizing Sentence-level Logical Document Structures with the Help of Context-free Grammars},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5282--5290},
  abstract = {Current sentence boundary detectors split documents into sequentially ordered sentences by detecting their beginnings and ends. Sentences, however, are more deeply structured even on this side of constituent and dependency structure: they can consist of a main sentence and several subordinate clauses as well as further segments (e.g. inserts in parentheses); they can even recursively embed whole sentences and then contain multiple sentence beginnings and ends. In this paper, we introduce a tool that segments sentences into tree structures to detect this type of recursive structure. To this end, we retrain different constituency parsers with the help of modified training data to transform them into sentence segmenters. With these segmenters, documents are mapped to sequences of sentence-related “logical document structures”. The resulting segmenters aim to improve downstream tasks by providing additional structural information. In this context, we experiment with German dependency parsing. We show that for certain sentence categories, which can be determined automatically, improvements in German dependency parsing can be achieved using our segmenter for preprocessing. The assumption

suggests that improvements in other languages and tasks can be achieved.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.650}
}

@InProceedings{guibon-EtAl:2020:LREC,
  author   = {Guibon, Gaël and Courtin, Marine and Gerdes, Kim and Guillaume, Bruno},
  title   = {When Collaborative Treebank Curation Meets Graph Grammars},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5291--5300},
  abstract = {In this paper we present Arborator-Grew, a collaborative annotation tool for treebank development. Arborator-Grew combines the features of two preexisting tools: Arborator and Grew. Arborator is a widely used collaborative graphical online dependency treebank annotation tool. Grew is a tool for graph querying and rewriting specialized in structures needed in NLP, i.e. syntactic and semantic dependency trees and graphs. Grew also has an online version, Grew-match, where all Universal Dependencies treebanks in their classical, deep and surface-syntactic flavors can be queried. Arborator-Grew is a complete redevelopment and modernization of Arborator, replacing its own internal database storage by a new Grew API, which adds a powerful query tool to Arborator's existing treebank creation and correction features. This includes complex access control for parallel expert and crowd-sourced annotation, tree comparison visualization, and various exercise modes for teaching and training of annotators. Arborator-Grew opens up new paths of collectively creating, updating, maintaining, and curating syntactic treebanks and semantic graph banks.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.651}
}

@InProceedings{wang-EtAl:2020:LREC2,
  author   = {Wang, Ilaine and Pelletier, Aurore and Antoine, Jean-Yves and Halftermeyer, Anaïs},
  title   = {ODIL_Syntax: a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5301--5307},
  abstract = {This paper describes ODIL Syntax, a French treebank built on spontaneous speech transcripts. The syntactic structure of every speech turn is represented by constituent trees, through a

procedure which combines an automatic annotation provided by a parser (here, the Stanford Parser) and a manual revision. ODIL Syntax respects the annotation scheme designed for the French TreeBank (FTB), with the addition of some annotation guidelines that aims at representing specific features of the spoken language such as speech disfluencies. The corpus will be freely distributed by January 2020 under a Creative Commons licence. It will ground a further semantic enrichment dedicated to the representation of temporal entities and temporal relations, as a second phase of the ODIL@Temporal project. The paper details the annotation scheme we followed with a emphasis on the representation of speech disfluencies. We then present the annotation procedure that was carried out on the Contemplata annotation platform. In the last section, we provide some distributional characteristics of the annotated corpus (POS distribution, multiword expressions).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.652}
}

@InProceedings{wrblewska:2020:LREC,
  author   = {Wróblewska, Alina},
  title   = {Towards the Conversion of National Corpus of Polish to Universal Dependencies},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5308--5315},
  abstract = {The research presented in this paper aims at enriching the manually morphosyntactically annotated part of National Corpus of Polish (NKJP1M) with a syntactic layer, i.e. dependency trees of sentences, and at converting both dependency trees and morphosyntactic annotations of particular tokens to Universal Dependencies. The dependency layer is built using a semi-automatic annotation procedure. The sentences from NKJP1M are first parsed with a dependency parser trained on Polish Dependency Bank, i.e. the largest bank of Polish dependency trees. The predicted dependency trees and the morphosyntactic annotations of tokens are then automatically converted into UD dependency graphs. NKJP1M sentences are an essential part of Polish Dependency Bank, we thus replace some automatically predicted dependency trees with their manually annotated equivalents. The final dependency treebank consists of 86K trees (including 15K gold-standard trees). A natural language pre-processing model trained on the enlarged set of (possibly noisy) dependency trees outperforms a model trained on a smaller set of the gold-standard trees in predicting part-of-speech tags, morphological features, lemmata, and labelled dependency trees},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.653}
}

@InProceedings{grossman-EtAl:2020:LREC,
  author   = {Grossman, Eitan and Eisen, Elad and Nikolaev,

Dmitry and Moran, Steven},
  title   = {SegBo: A Database of Borrowed Sounds in the World's Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5316--5322},
  abstract = {Phonological segment borrowing is a process through which languages acquire new contrastive speech sounds as the result of borrowing new words from other languages. Despite the fact that phonological segment borrowing is documented in many of the world's languages, to date there has been no large-scale quantitative study of the phenomenon. In this paper, we present SegBo, a novel cross-linguistic database of borrowed phonological segments. We describe our data aggregation pipeline and the resulting language sample. We also present two short case studies based on the database. The first deals with the impact of large colonial languages on the sound systems of the world's languages; the second deals with universals of borrowing in the domain of rhotic consonants.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.654}
}

@InProceedings{lancien-ct-bigi:2020:LREC,
  author   = {Lancien, Mélanie and Côté, Marie-Hélène and Bigi, Brigitte},
  title   = {Developing Resources for Automated Speech Processing of Quebec French},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5323--5328},
  abstract = {The analysis of the structure of speech nearly always rests on the alignment of the speech recording with a phonetic transcription. Nowadays several tools can perform this speech segmentation automatically. However, none of them allows the automatic segmentation of Quebec French (QF hereafter), the acoustics and phonotactics of QF differing widely from that of France French (FF hereafter). To adequately segment QF, features like diphthongization of long vowels and affrication of coronal stops have to be taken into account. Thus acoustic models for automatic segmentation must be trained on speech samples exhibiting those phenomena. Dictionaries and lexicons must also be adapted and integrate differences in lexical units and in the phonology of QF. This paper presents the development of linguistic resources to be included into SPPAS software tool in order to get Text normalization, Phonetization, Alignment and Syllabification. We adapted the existing French lexicon and developed a QF-specific pronunciation dictionary. We then created an acoustic model from the

```
existing ones and adapted it with 5 minutes of manually time-aligned
data. These new resources are all freely distributed with SPPAS
version 2.7; they perform the full process of speech segmentation in
Quebec French.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.655}
}

@InProceedings{mortensen-EtAl:2020:LREC,
  author    = {Mortensen, David R.  and  Li, Xinjian  and  Littell,
Patrick  and  Michaud, Alexis  and  Rijhwani, Shruti  and
Anastasopoulos, Antonios  and  Black, Alan W  and  Metze, Florian
and  Neubig, Graham},
  title     = {AlloVera: A Multilingual Allophone Database},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5329--5336},
  abstract  = {We introduce a new resource, AlloVera, which provides
mappings from 218 allophones to phonemes for 14 languages. Phonemes
are contrastive phonological units, and allophones are their various
concrete realizations, which are predictable from phonological
context. While phonemic representations are language specific,
phonetic representations (stated in terms of (allo)phones) are much
closer to a universal (language-independent) transcription. AlloVera
allows the training of speech recognition models that output
phonetic transcriptions in the International Phonetic Alphabet
(IPA), regardless of the input language. We show that a “universal”
allophone model, Allosaurus, built with AlloVera, outperforms
“universal” phonemic models and language-specific models on a
speech-transcription task. We explore the implications of this
technology (and related technologies) for the documentation of
endangered and minority languages. We further explore other
applications for which AlloVera will be suitable as it grows,
including phonological typology.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.656}
}

@InProceedings{ibrahim-EtAl:2020:LREC,
  author    = {Ibrahim, Omnia  and  Asadi, Homa  and  Kassem, Eman
and  Dellwo, Volker},
  title     = {Arabic Speech Rhythm Corpus: Read and Spontaneous
Speaking Styles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5337--5342},
  abstract  = {Databases for studying speech rhythm and tempo exist
for numerous languages. The present corpus was built to allow
```

```
comparisons between Arabic speech rhythm and other languages. 10
Egyptian speakers (gender-balanced) produced speech in two different
speaking styles (read and spontaneous). The design of the reading
task replicates the methodology used in the creation of BonnTempo
corpus (BTC). During the spontaneous task, speakers talked freely
for more than one minute about their daily life and/or their
studies, then they described the directions to come to the
university from a famous near location using a map as a visual
stimulus. For corpus annotation, the database has been manually and
automatically time-labeled, which makes it feasible to perform a
quantitative analysis of the rhythm of Arabic in both Modern
Standard Arabic (MSA) and Egyptian dialect variety. The database
serves as a phonetic resource, which allows researchers to examine
various aspects of Arabic supra-segmental features and it can be
used for forensic phonetic research, for comparison of different
speakers, analyzing variability in different speaking styles, and
automatic speech and speaker recognition.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.657}
}

@InProceedings{johannessen-EtAl:2020:LREC,
  author    = {Johannessen, Janne  and  Kåsen, Andre  and  Hagen,
Kristin  and  Nøklestad, Anders  and  Priestley, Joel},
  title     = {Comparing Methods for Measuring Dialect Similarity in
Norwegian},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5343--5350},
  abstract  = {The present article presents four experiments with
two different methods for measuring dialect similarity in Norwegian:
the Levenshtein method and the neural long short term memory (LSTM)
autoencoder network, a machine learning algorithm. The visual output
in the form of dialect maps is then compared with canonical maps
found in the dialect literature. All of this enables us to say that
one does not need fine-grained transcriptions of speech to replicate
classical classification patterns.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.658}
}

@InProceedings{ahamad-anand-bhargava:2020:LREC,
  author    = {Ahamad, Afroz  and  Anand, Ankit  and  Bhargava,
Pranesh},
  title     = {AccentDB: A Database of Non-Native English Accents to
Assist Neural Speech Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {5351--5358},
  abstract  = {Modern Automatic Speech Recognition (ASR) technology
has evolved to identify the speech spoken by native speakers of a
language very well. However, identification of the speech spoken by
non-native speakers continues to be a major challenge for it. In
this work, we first spell out the key requirements for creating a
well-curated database of speech samples in non-native accents for
training and testing robust ASR systems. We then introduce AccentDB,
one such database that contains samples of 4 Indian-English accents
collected by us, and a compilation of samples from 4 native-English,
and a metropolitan Indian-English accent. We also present an
analysis on separability of the collected accent data. Further, we
present several accent classification models and evaluate them
thoroughly against human-labelled accent classes. We test the
generalization of our classifier models in a variety of setups of
seen and unseen data. Finally, we introduce accent neutralization of
non-native accents to native accents using autoencoder models with
task-specific architectures. Thus, our work aims to aid ASR systems
at every stage of development with a database for training,
classification models for feature augmentation, and neutralization
systems for acoustic transformations of non-native accents of
English.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.659}
}

@InProceedings{zarcone-alam-kolagar:2020:LREC,
  author    = {Zarcone, Alessandra  and  Alam, Touhidul  and
Kolagar, Zahra},
  title     = {PATE: A Corpus of Temporal Expressions for the In-car
Voice Assistant Domain},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {523--530},
  abstract  = {The recognition and automatic annotation of temporal
expressions (e.g. "Add an event for tomorrow evening at eight to my
calendar") is a key module for AI voice assistants, in order to
allow them to interact with apps (for example, a calendar app).
However, in the NLP literature, research on temporal expressions has
focused mostly on data from the news, from the clinical domain, and
from social media. The voice assistant domain is very different than
the typical domains that have been the focus of work on temporal
expression identification, thus requiring a dedicated data
collection. We present a crowdsourcing method for eliciting natural-
language commands containing temporal expressions for an AI voice
assistant, by using pictures and scenario descriptions. We annotated
the elicited commands (480) as well as the commands in the Snips
dataset following the TimeML/TIMEX3 annotation guidelines, reaching
a total of 1188 annotated commands. The commands can be later used
to train the NLU components of an AI voice assistant.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.66}
```

}

@InProceedings{schlegel-EtAl:2020:LREC,
  author   = {Schlegel, Viktor and Valentino, Marco and Freitas, Andre and Nenadic, Goran and Batista-Navarro, Riza},
  title   = {A Framework for Evaluation of Machine Reading Comprehension Gold Standards},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5359--5369},
  abstract = {Machine Reading Comprehension (MRC) is the task of answering a question over a paragraph of text. While neural MRC systems gain popularity and achieve noticeable performance, issues are being raised with the methodology used to establish their performance, particularly concerning the data design of gold standards that are used to evaluate them. There is but a limited understanding of the challenges present in this data, which makes it hard to draw comparisons and formulate reliable hypotheses. As a first step towards alleviating the problem, this paper proposes a unifying framework to systematically investigate the present linguistic features, required reasoning and background knowledge and factual correctness on one hand, and the presence of lexical cues as a lower bound for the requirement of understanding on the other hand. We propose a qualitative annotation schema for the first and a set of approximative metrics for the latter. In a first application of the framework, we analyse modern MRC gold standards and present our findings: the absence of features that contribute towards lexical ambiguity, the varying factual correctness of the expected answers and the presence of lexical cues, all of which potentially lower the reading comprehension complexity and quality of the evaluation data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.660}
}

@InProceedings{xu-EtAl:2020:LREC2,
  author   = {Xu, Dongfang and Jansen, Peter and Martin, Jaycie and Xie, Zhengnan and Yadav, Vikas and Tayyar Madabushi, Harish and Tafjord, Oyvind and Clark, Peter},
  title   = {Multi-class Hierarchical Question Classification for Multiple Choice Science Exams},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5370--5382},
  abstract = {Prior work has demonstrated that question classification (QC), recognizing the problem domain of a question, can help answer it more accurately. However, developing strong QC

```
algorithms has been hindered by the limited size and complexity of
annotated data available. To address this, we present the largest
challenge dataset for QC, containing 7,787 science exam questions
paired with detailed classification labels from a fine-grained
hierarchical taxonomy of 406 problem domains. We then show that a
BERT-based model trained on this dataset achieves a large (+0.12
MAP) gain compared with previous methods, while also achieving
state-of-the-art performance on benchmark open-domain and biomedical
QC datasets. Finally, we show that using this model's predictions of
question topic significantly improves the accuracy of a question
answering system by +1.7\% P@1, with substantial future gains
possible as QC performance improves.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.661}
}

@InProceedings{woldemariam:2020:LREC,
  author    = {Woldemariam, Yonas},
  title     = {Assessing Users' Reputation from Syntactic and
Semantic Information in Community Question Answering},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5383--5391},
  abstract  = {Textual content is the most significant as well as
substantially the big part of CQA (Community Question Answering)
forums. Users gain reputation for contributing such content.
Although linguistic quality is the very essence of textual
information, that does not seem to be considered in estimating
users' reputation. As existing users' reputation systems seem to
solely rely on vote counting, adding that bit of linguistic
information surely improves their quality. In this study, we
investigate the relationship between users' reputation and
linguistic features extracted from their associated answers content.
And we build statistical models on a Stack Overflow dataset that
learn reputation from complex syntactic and semantic structures of
such content. The resulting models reveal how users' writing styles
in answering questions play important roles in building reputation
points. In our experiments, extracting answers from systematically
selected users followed by linguistic features annotation and models
building. The models are evaluated on in-domain (e.g., Server Fault,
Super User) and out-domain (e.g., English, Maths) datasets. We found
out that the selected linguistic features have quite significant
influences over reputation scores. In the best case scenario, the
selected linguistic feature set could explain 80\% variation in
reputation scores with the prediction error of 3\%. The performance
results obtained from the baseline models have been significantly
improved by adding syntactic and punctuation marks features.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.662}
}

@InProceedings{nishida-EtAl:2020:LREC,
```

```
  author    = {Nishida, Kosuke  and  Nishida, Kyosuke  and  Saito,
Itsumi  and  Asano, Hisako  and  Tomita, Junji},
  title     = {Unsupervised Domain Adaptation of Language Models for
Reading Comprehension},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5392--5399},
  abstract  = {This study tackles unsupervised domain adaptation of
reading comprehension (UDARC). Reading comprehension (RC) is a task
to learn the capability for question answering with textual sources.
State-of-the-art models on RC still do not have general linguistic
intelligence; i.e., their accuracy worsens for out-domain datasets
that are not used in the training. We hypothesize that this
discrepancy is caused by a lack of the language modeling (LM)
capability for the out-domain. The UDARC task allows models to use
supervised RC training data in the source domain and only unlabeled
passages in the target domain. To solve the UDARC problem, we
provide two domain adaptation models. The first one learns the out-
domain LM and in-domain RC task sequentially. The second one is the
proposed model that uses a multi-task learning approach of LM and
RC. The models can retain both the RC capability acquired from the
supervised data in the source domain and the LM capability from the
unlabeled data in the target domain. We evaluated the models on
UDARC with five datasets in different domains. The models
outperformed the model without domain adaptation. In particular, the
proposed model yielded an improvement of 4.3/4.2 points in EM/F1 in
an unseen biomedical domain.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.663}
}

@InProceedings{yoon-EtAl:2020:LREC,
  author    = {Yoon, Seunghyun  and  Dernoncourt, Franck  and  Kim,
Doo Soon  and  Bui, Trung  and  Jung, Kyomin},
  title     = {Propagate-Selector: Detecting Supporting Sentences
for Question Answering via Graph Neural Networks},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5400--5407},
  abstract  = {In this study, we propose a novel graph neural
network called propagate-selector (PS), which propagates information
over sentences to understand information that cannot be inferred
when considering sentences in isolation. First, we design a graph
structure in which each node represents an individual sentence, and
some pairs of nodes are selectively connected based on the text
structure. Then, we develop an iterative attentive aggregation and a
skip-combine method in which a node interacts with its neighborhood
```

nodes to accumulate the necessary information. To evaluate the performance of the proposed approaches, we conduct experiments with the standard HotpotQA dataset. The empirical results demonstrate the superiority of our proposed approach, which obtains the best performances, compared to the widely used answer-selection models that do not consider the intersentential relationship.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.664}
}

@InProceedings{cortes-EtAl:2020:LREC,
  author   = {Cortes, Eduardo and Woloszyn, Vinicius and Binder, Arne and Himmelsbach, Tilo and Barone, Dante and Möller, Sebastian},
  title   = {An Empirical Comparison of Question Classification Methods for Question Answering Systems},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5408--5416},
  abstract = {Question classification is an important component of Question Answering Systems responsible for identifying the type of an answer a particular question requires. For instance, ``Who is the prime minister of the United Kingdom?" demands a name of a PERSON, while ``When was the queen of the United Kingdom born?" entails a DATE. This work makes an extensible review of the most recent methods for Question Classification, taking into consideration their applicability in low-resourced languages. First, we propose a manual classification of the current state-of-the-art methods in four distinct categories: low, medium, high, and very high level of dependency on external resources. Second, we applied this categorization in an empirical comparison in terms of the amount of data necessary for training and performance in different languages. In addition to complementing earlier works in this field, our study shows a boost on methods relying on recent language models, overcoming methods not suitable for low-resourced languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.665}
}

@InProceedings{wu-hao:2020:LREC,
  author   = {Wu, Jinmeng and Hao, Yanbin},
  title   = {Cross-sentence Pre-trained Model for Interactive QA matching},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5417--5424},
  abstract = {Semantic matching measures the dependencies between query and answer representations, it is an important criterion for

evaluating whether the matching is successful. In fact, such matching does not examine each sentence individually, context information outside a sentence should be considered equally important to the syntactic context inside a sentence. We proposed a new QA matching model, built upon a cross-sentence context-aware architecture. An interactive attention mechanism with a pre-trained language model is proposed to automatically select salient positional answer representations that contribute more significantly to the answer relevance of a given question. In addition to the context information captured at each word position, we incorporate a new quantity of context information jump to facilitate the attention weight formulation. This reflects the amount of new information brought by the next word and is computed by modeling the joint probability between two adjacent word states. The proposed method is compared to multiple state-of-the-art ones evaluated using the TREC library, WikiQA, and the Yahoo! community question datasets. Experimental results show that the proposed method outperforms satisfactorily the competing ones.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.666}
}

@InProceedings{lee-hwang-cho:2020:LREC,
  author   = {Lee, Gyeongbok and Hwang, Seung-won and Cho, Hyunsouk},
  title   = {SQuAD2-CR: Semi-supervised Annotation for Cause and Rationales for Unanswerability in SQuAD 2.0},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5425--5432},
  abstract = {Existing machine reading comprehension models are reported to be brittle for adversarially perturbed questions when optimizing only for accuracy, which led to the creation of new reading comprehension benchmarks, such as SQuAD 2.0 which contains such type of questions. However, despite the super-human accuracy of existing models on such datasets, it is still unclear how the model predicts the answerability of the question, potentially due to the absence of a shared annotation for the explanation. To address such absence, we release SQuAD2-CR dataset, which contains annotations on unanswerable questions from the SQuAD 2.0 dataset, to enable an explanatory analysis of the model prediction. Specifically, we annotate (1) explanation on why the most plausible answer span cannot be the answer and (2) which part of the question causes unanswerability. We share intuitions and experimental results that how this dataset can be used to analyze and improve the interpretability of existing reading comprehension model behavior.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.667}
}

@InProceedings{kodama-EtAl:2020:LREC,
  author   = {Kodama, Takashi and Higashinaka, Ryuichiro and

Mitsuda, Koh and Masumura, Ryo and Aono, Yushi and Nakamura, Ryuta and Adachi, Noritake and Kawabata, Hidetoshi},
  title   = {Generating Responses that Reflect Meta Information in User-Generated Question Answer Pairs},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5433--5441},
  abstract = {This paper concerns the problem of realizing consistent personalities in neural conversational modeling by using user generated question-answer pairs as training data. Using the framework of role play-based question answering, we collected single-turn question-answer pairs for particular characters from online users. Meta information was also collected such as emotion and intimacy related to question-answer pairs. We verified the quality of the collected data and, by subjective evaluation, we also verified their usefulness in training neural conversational models for generating utterances reflecting the meta information, especially emotion.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.668}
}

@InProceedings{gonalooliveira-EtAl:2020:LREC,
  author   = {Gonçalo Oliveira, Hugo and Ferreira, João and Santos, José and Fialho, Pedro and Rodrigues, Ricardo and Coheur, Luisa and Alves, Ana},
  title   = {AIA-BDE: A Corpus of FAQs in Portuguese and their Variations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5442--5449},
  abstract = {We present AIA-BDE, a corpus of 380 domain-oriented FAQs in Portuguese and their variations, i.e., paraphrases or entailed questions, created manually, by humans, or automatically, with Google Translate. Its aims to be used as a benchmark for FAQ retrieval and automatic question-answering, but may be useful in other contexts, such as the development of task-oriented dialogue systems, or models for natural language inference in an interrogative context. We also report on two experiments. Matching variations with their original questions was not trivial with a set of unsupervised baselines, especially for manually created variations. Besides high performances obtained with ELMo and BERT embeddings, an Information Retrieval system was surprisingly competitive when considering only the first hit. In the second experiment, text classifiers were trained with the original questions, and tested when assigning each variation to one of three possible sources, or assigning them as out-of-domain. Here, the

difference between manual and automatic variations was not so significant.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.669}
}

@InProceedings{ribeiro-ribeiro-martinsdematos:2020:LREC,
  author   = {Ribeiro, Eugénio and Ribeiro, Ricardo and Martins de Matos, David},
  title   = {Mapping the Dialog Act Annotations of the LEGO Corpus into ISO 24617-2 Communicative Functions},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {531--539},
  abstract = {ISO 24617-2, the ISO standard for dialog act annotation, sets the ground for more comparable research in the area. However, the amount of data annotated according to it is still reduced, which impairs the development of approaches for automatic recognition. In this paper, we describe a mapping of the original dialog act labels of the LEGO corpus, which have been neglected, into the communicative functions of the standard. Although this does not lead to a complete annotation according to the standard, the 347 dialogs provide a relevant amount of data that can be used in the development of automatic communicative function recognition approaches, which may lead to a wider adoption of the standard. Using the 17 English dialogs of the DialogBank as gold standard, our preliminary experiments have shown that including the mapped dialogs during the training phase leads to improved performance while recognizing communicative functions in the Task dimension.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.67}
}

@InProceedings{colas-EtAl:2020:LREC,
  author   = {Colas, Anthony and Kim, Seokhwan and Dernoncourt, Franck and Gupte, Siddhesh and Wang, Zhe and Kim, Doo Soon},
  title   = {TutorialVQA: Question Answering Dataset for Tutorial Videos},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5450--5455},
  abstract = {Despite the number of currently available datasets on video-question answering, there still remains a need for a dataset involving multi-step and non-factoid answers. Moreover, relying on video transcripts remains an under-explored topic. To adequately address this, we propose a new question answering task on instructional videos, because of their verbose and narrative nature. While previous studies on video question answering have focused on

generating a short text as an answer, given a question and video clip, our task aims to identify a span of a video segment as an answer which contains instructional details with various granularities. This work focuses on screencast tutorial videos pertaining to an image editing program. We introduce a dataset, TutorialVQA, consisting of about 6,000 manually collected triples of (video, question, answer span). We also provide experimental results with several baseline algorithms using the video transcripts. The results indicate that the task is challenging and call for the investigation of new algorithms.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.670}
}

@InProceedings{xie-EtAl:2020:LREC,
  author   = {Xie, Zhengnan and Thiem, Sebastian and Martin, Jaycie and Wainwright, Elizabeth and Marmorstein, Steven and Jansen, Peter},
  title   = {WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5456--5473},
  abstract = {Explainable question answering for complex questions often requires combining large numbers of facts to answer a question while providing a human-readable explanation for the answer, a process known as multi-hop inference. Standardized science questions require combining an average of 6 facts, and as many as 16 facts, in order to answer and explain, but most existing datasets for multi-hop reasoning focus on combining only two facts, significantly limiting the ability of multi-hop inference algorithms to learn to generate large inferences. In this work we present the second iteration of the WorldTree project, a corpus of 5,114 standardized science exam questions paired with large detailed multi-fact explanations that combine core scientific knowledge and world knowledge. Each explanation is represented as a lexically-connected "explanation graph" that combines an average of 6 facts drawn from a semi-structured knowledge base of 9,216 facts across 66 tables. We use this explanation corpus to author a set of 344 high-level science domain inference patterns similar to semantic frames supporting multi-hop inference. Together, these resources provide training data and instrumentation for developing many-fact multi-hop inference models for question answering.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.671}
}

@InProceedings{luthier-popescubelis:2020:LREC,
  author   = {Luthier, Gabriel and Popescu-Belis, Andrei},
  title   = {Chat or Learn: a Data-Driven Robust Question-Answering System},
  booktitle   = {Proceedings of The 12th Language Resources and

```
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5474--5480},
  abstract  = {We present a voice-based conversational agent which
combines the robustness of chatbots and the utility of question
answering (QA) systems. Indeed, while data-driven chatbots are
typically user-friendly but not goal-oriented, QA systems tend to
perform poorly at chitchat. The proposed chatbot relies on a
controller which performs dialogue act classification and feeds user
input either to a sequence-to-sequence chatbot or to a QA system.
The resulting chatbot is a spoken QA application for the Google Home
smart speaker. The system is endowed with general-domain knowledge
from Wikipedia articles and uses coreference resolution to detect
relatedness between questions. We present our choices of data sets
for training and testing the components, and present the
experimental results that helped us optimize the parameters of the
chatbot. In particular, we discuss the appropriateness of using the
SQuAD dataset for evaluating end-to-end QA, in the light of our
system's behavior.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.672}
}

@InProceedings{keraron-EtAl:2020:LREC,
  author    = {Keraron, Rachel  and  Lancrenon, Guillaume  and
Bras, Mathilde  and  Allary, Frédéric  and  Moyse, Gilles  and
Scialom, Thomas  and  Soriano-Morales, Edmundo-Pavel  and  Staiano,
Jacopo},
  title     = {Project PIAF: Building a Native French Question-
Answering Dataset},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5481--5490},
  abstract  = {Motivated by the lack of data for non-English
languages, in particular for the evaluation of downstream tasks such
as Question Answering, we present a participatory effort to collect
a native French Question Answering Dataset. Furthermore, we describe
and publicly release the annotation tool developed for our
collection effort, along with the data obtained and preliminary
baselines.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.673}
}

@InProceedings{charlet-EtAl:2020:LREC,
  author    = {Charlet, Delphine  and  Damnati, Geraldine  and
Bechet, Frederic  and  marzinotto, gabriel  and  Heinecke,
Johannes},
  title     = {Cross-lingual and Cross-domain Evaluation of Machine
```

Reading Comprehension with Squad and CALOR-Quest Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5491--5497},
  abstract = {Machine Reading received recently a lot of attention thanks to both the availability of very large corpora such as SQuAD or MS MARCO containing triplets (document, question, answer), and the introduction of Transformer Language Models such as BERT which obtain excellent results, even matching human performance according to the SQuAD leaderboard. One of the key features of Transformer Models is their ability to be jointly trained across multiple languages, using a shared subword vocabulary, leading to the construction of cross-lingual lexical representations. This feature has been used recently to perform zero-shot cross-lingual experiments where a multilingual BERT model fine-tuned on a machine reading comprehension task exclusively for English was directly applied to Chinese and French documents with interesting performance. In this paper we study the cross-language and cross-domain capabilities of BERT on a Machine Reading Comprehension task on two corpora: SQuAD and a new French Machine Reading dataset, called CALOR-QUEST. The semantic annotation available on CALOR-QUEST allows us to give a detailed analysis on the kinds of questions that are properly handled through the cross-language process. We will try to answer this question: which factor between language mismatch and domain mismatch has the strongest influence on the performances of a Machine Reading Comprehension task?},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.674}
}

@InProceedings{saikh-ekbal-bhattacharyya:2020:LREC,
  author   = {Saikh, Tanik and Ekbal, Asif and Bhattacharyya, Pushpak},
  title   = {ScholarlyRead: A New Dataset for Scientific Article Reading Comprehension},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5498--5504},
  abstract = {We present ScholarlyRead, span-of-word-based scholarly articles' Reading Comprehension (RC) dataset with approximately 10K manually checked passage-question-answer instances. ScholarlyRead was constructed in semi-automatic way. We consider the articles from two popular journals of a reputed publishing house. Firstly, we generate questions from these articles in an automatic way. Generated questions are then manually checked by the human annotators. We propose a baseline model based on Bi-Directional Attention Flow (BiDAF) network that yields the F1 score

of 37.31\%. The framework would be useful for building Question-Answering (QA) systems on scientific articles.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.675}
}

@InProceedings{laskar-huang-hoque:2020:LREC,
  author   = {Laskar, Md Tahmid Rahman and Huang, Jimmy Xiangji and Hoque, Enamul},
  title   = {Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5505--5514},
  abstract = {Word embeddings that consider context have attracted great attention for various natural language processing tasks in recent years. In this paper, we utilize contextualized word embeddings with the transformer encoder for sentence similarity modeling in the answer selection task. We present two different approaches (feature-based and fine-tuning-based) for answer selection. In the feature-based approach, we utilize two types of contextualized embeddings, namely the Embeddings from Language Models (ELMo) and the Bidirectional Encoder Representations from Transformers (BERT) and integrate each of them with the transformer encoder. We find that integrating these contextual embeddings with the transformer encoder is effective to improve the performance of sentence similarity modeling. In the second approach, we fine-tune two pre-trained transformer encoder models for the answer selection task. Based on our experiments on six datasets, we find that the fine-tuning approach outperforms the feature-based approach on all of them. Among our fine-tuning-based models, the Robustly Optimized BERT Pretraining Approach (RoBERTa) model results in new state-of-the-art performance across five datasets.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.676}
}

@InProceedings{carrino-costajuss-fonollosa:2020:LREC,
  author   = {Carrino, Casimiro Pio and Costa-jussà, Marta R. and Fonollosa, José A. R.},
  title   = {Automatic Spanish Translation of SQuAD Dataset for Multi-lingual Question Answering},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5515--5523},
  abstract = {Recently, multilingual question answering became a crucial research topic, and it is receiving increased interest in the NLP community. However, the unavailability of large-scale

datasets makes it challenging to train multilingual QA systems with performance comparable to the English ones. In this work, we develop the Translate Align Retrieve (TAR) method to automatically translate the Stanford Question Answering Dataset (SQuAD) v1.1 to Spanish. We then used this dataset to train Spanish QA systems by fine-tuning a Multilingual-BERT model. Finally, we evaluated our QA models with the recently proposed MLQA and XQuAD benchmarks for cross-lingual Extractive QA. Experimental results show that our models outperform the previous Multilingual-BERT baselines achieving the new state-of-the-art values of 68.1 F1 on the Spanish MLQA corpus and 77.6 F1 on the Spanish XQuAD corpus. The resulting, synthetically generated SQuAD-es v1.1 corpora, with almost 100\% of data contained in the original English version, to the best of our knowledge, is the first large-scale QA training resource for Spanish.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.677}
}

@InProceedings{alizadeh-dieugenio:2020:LREC,
  author   = {Alizadeh, Mehrdad and Di Eugenio, Barbara},
  title   = {A Corpus for Visual Question Answering Annotated with Frame Semantic Information},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5524--5531},
  abstract = {Visual Question Answering (VQA) has been widely explored as a computer vision problem, however enhancing VQA systems with linguistic information is necessary for tackling the complexity of the task. The language understanding part can play a major role especially for questions asking about events or actions expressed via verbs. We hypothesize that if the question focuses on events described by verbs, then the model should be aware of or trained with verb semantics, as expressed via semantic role labels, argument types, and/or frame elements. Unfortunately, no VQA dataset exists that includes verb semantic information. We created a new VQA dataset annotated with verb semantic information called imSituVQA. imSituVQA is built by taking advantage of the imSitu dataset annotations. The imSitu dataset consists of images manually labeled with semantic frame elements, mostly taken from FrameNet.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.678}
}

@InProceedings{soni-roberts:2020:LREC,
  author   = {Soni, Sarvesh and Roberts, Kirk},
  title   = {Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5532--5538},
  abstract  = {We evaluate the performance of various Transformer
language models, when pre-trained and fine-tuned on different
combinations of open-domain, biomedical, and clinical corpora on two
clinical question answering (QA) datasets (CliCR and emrQA). We
perform our evaluations on the task of machine reading
comprehension, which involves training the model to answer a
question given an unstructured context paragraph. We conduct a total
of 48 experiments on different combinations of the large open-domain
and domain-specific corpora. We found that an initial fine-tuning on
an open-domain dataset, SQuAD, consistently improves the clinical QA
performance across all the model variants.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.679}
}

@InProceedings{miehle-EtAl:2020:LREC,
  author    = {Miehle, Juliana  and  Feustel, Isabel  and  Hornauer,
Julia  and  Minker, Wolfgang  and  Ultes, Stefan},
  title     = {Estimating User Communication Styles for Spoken
Dialogue Systems},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {540--548},
  abstract  = {We present a neural network approach to estimate the
communication style of spoken interaction, namely the stylistic
variations elaborateness and directness, and investigate which type
of input features to the estimator are necessary to achive good
performance. First, we describe our annotated corpus of recordings
in the health care domain and analyse the corpus statistics in terms
of agreement, correlation and reliability of the ratings. We use
this corpus to estimate the elaborateness and the directness of each
utterance. We test different feature sets consisting of dialogue act
features, grammatical features and linguistic features as input for
our classifier and perform classification in two and three classes.
Our classifiers use only features that can be automatically derived
during an ongoing interaction in any spoken dialogue system without
any prior annotation. Our results show that the elaborateness can be
classified by only using the dialogue act and the amount of words
contained in the corresponding utterance. The directness is a more
difficult classification task and additional linguistic features in
form of word embeddings improve the classification results.
Afterwards, we run a comparison with a support vector machine and a
recurrent neural network classifier.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.68}
}

@InProceedings{branco-EtAl:2020:LREC2,
  author    = {Branco, António  and  Calzolari, Nicoletta  and
```

Vossen, Piek and Van Noord, Gertjan and van Uytvanck, Dieter and Silva, João and Gomes, Luís and Moreira, André and Elbers, Willem},
  title   = {A Shared Task of a New, Collaborative Type to Foster Reproducibility: A First Exercise in the Area of Language Science and Technology with REPROLANG2020},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5539--5545},
  abstract = {n this paper, we introduce a new type of shared task – which is collaborative rather than competitive – designed to support and fosterthe reproduction of research results. We also describe the first event running such a novel challenge, present the results obtained, discussthe lessons learned and ponder on future undertakings.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.680}
}

@InProceedings{garneau-EtAl:2020:LREC,
  author   = {Garneau, Nicolas and Godbout, Mathieu and Beauchemin, David and Durand, Audrey and Lamontagne, Luc},
  title   = {A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings: Making the Method Robustly Reproducible as Well},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5546--5554},
  abstract = {In this paper, we reproduce the experiments of Artetxe et al. (2018b) regarding the robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. We show that the reproduction of their method is indeed feasible with some minor assumptions. We further investigate the robustness of their model by introducing four new languages that are less similar to English than the ones proposed by the original paper. In order to assess the stability of their model, we also conduct a grid search over sensible hyperparameters. We then propose key recommendations that apply to any research project in order to deliver fully reproducible research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.681}
}

@InProceedings{pluciski-lango-zimniewicz:2020:LREC,
  author   = {Plucíński, Kamil and Lango, Mateusz and Zimniewicz, Michał},
  title   = {A Closer Look on Unsupervised Cross-lingual Word Embeddings Mapping},

  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5555--5562},
  abstract = {In this work, we study the unsupervised cross-lingual word embeddings mapping method presented by Artetxe et al. (2018). First, wesuccessfully reproduced the experiments performed in the original work, finding only minor differences. Furthermore, we verified themethod’s robustness on different embedding representations and new language pairs, particularly these involving Slavic languages likePolish or Czech. We also performed an experimental analysis of the impact of the method’s parameters on the final result. Finally, welooked for an alternative way of initialization, which directly relies on the isometric assumption. Our work confirms the results presentedearlier, at the same time pointing at interesting problems occurring while using the method with different types of embeddings or onless-common language pairs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.682}
}

@InProceedings{khoe:2020:LREC,
  author   = {Khoe, Yung Han},
  title   = {Reproducing a Morphosyntactic Tagger with a Meta-BiLSTM Model over Context Sensitive Token Encodings},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5563--5568},
  abstract = {Reproducibility is generally regarded as being a requirement for any form of experimental science. Even so, reproduction of research results is only recently beginning to be practiced and acknowledged. In the context of the REPROLANG 2020 shared task, we contribute to this trend by reproducing the work reported on by Bohnet et al. (2018) on morphosyntactic tagging. Their meta-BiLSTM model achieved state-of-the-art results across a wide range of languages. This was done by integrating sentence-level and single-word context through synchronized training by a meta-model. Our reproduction only partially confirms the main results of the paper in terms of outperforming earlier models. The results of our reproductions improve on earlier models on the morphological tagging task, but not on the part-of-speech tagging task. Furthermore, even where we improve on earlier models, we fail to match the F1-scores reported for the meta-BiLSTM model. Because we chose not to contact the original authors for our reproduction study, the uncertainty about the degree of parallelism that was achieved between the original study and our reproduction limits the value of our findings as an assessment of the reliability of the

original results. At the same time, however, it underscores the relevance of our reproduction effort in regard to the reproducibility and interpretability of those findings. The discrepancies between our findings and the original results demonstrate that there is room for improvement in many aspects of reporting regarding the reproducibility of the experiments. In addition, we suggest that different reporting choices could improve the interpretability of the results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.683}
}

@InProceedings{rim-EtAl:2020:LREC1,
  author   = {Rim, Kyeongmin and Tu, Jingxuan and Lynch, Kelley and Pustejovsky, James},
  title   = {Reproducing Neural Ensemble Classifier for Semantic Relation Extraction inScientific Papers},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5569--5578},
  abstract = {Within the natural language processing (NLP) community, shared tasks play an important role. They define a common goal and allowthe the comparison of different methods on the same data. SemEval-2018 Task 7 involves the identification and classification of relationsin abstracts from computational linguistics (CL) publications. In this paper we describe an attempt to reproduce the methods and resultsfrom the top performing system at for SemEval-2018 Task 7. We describe challenges we encountered in the process, report on the resultsof our system, and discuss the ways that our attempt at reproduction can inform best practices.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.684}
}

@InProceedings{abdellatif-elgammal:2020:LREC,
  author   = {Abdellatif, Mohamed and Elgammal, Ahmed},
  title   = {ULMFiT replication},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5579--5587},
  abstract = {Authors: Mohamed Abdellatif and Ahmed Elgammal Gitlab URL: https://gitlab.com/abdollatif/lrec_app Commit hash: 3f20b2ddb96d8c865e5f56f5566edf371214785f Tag name: Splits2 Dataset file md5: 5aee3dac5e48d1ac3d279083212734c9 Dataset URL: https://drive.google.com/file/d/1cv5HuQhgFVizupFI40dzreemS2gMM498/view?usp=sharing},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.685}
}

```
@InProceedings{cooper-shardlow:2020:LREC,
  author    = {Cooper, Michael  and  Shardlow, Matthew},
  title     = {CombiNMT: An Exploration into Neural Text
Simplification Models},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5588--5594},
  abstract  = {This work presents a replication study of Exploring
Neural Text Simplification Models (Nisioi et al., 2017). We were
able to successfully replicate and extend the methods presented in
the original paper. Alongside the replication results, we present
our improvements dubbed CombiNMT. By using an updated implementation
of OpenNMT, and incorporating the Newsela corpus alongside the
original Wikipedia dataset (Hwang et al., 2016), as well as refining
both datasets to select high quality training examples. Our work
present two new systems, CombiNMT995, which is a result of matched
sentences with a cosine similarity of 0.995 or less, and CombiNMT98,
which, similarly, runs on a cosine similarity of 0.98 or less. By
extending the human evaluation presented within the original paper,
increasing both the number of annotators and the number of sentences
annotated, with the intention of increasing the quality of the
results, CombiNMT998 shows significant improvement over any of the
Neural Text Simplification (NTS) systems from the original paper in
terms of both the number of changes and the percentage of correct
changes made.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.686}
}

@InProceedings{bestgen:2020:LREC,
  author    = {Bestgen, Yves},
  title     = {Reproducing Monolingual, Multilingual and Cross-
Lingual CEFR Predictions},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5595--5602},
  abstract  = {his study aims to reproduce the research of Vajjala
and Rama (2018) which showed that it is possible to predict the
quality of a text written by learners of a given language by means
of a model built on the basis of texts written by learners of
another language. These authors also pointed out that POStag and
dependency n-grams were significantly more effective than text
length and global linguistic indices frequently used for this kind
of task. The analyses performed show that some important points of
their code did not correspond to the explanations given in the
paper. These analyses confirm the possibility to use syntactic n-
```

gram features in cross-lingual experiments to categorize texts according to their CEFR level (Common European Framework of Reference for Languages). However, text length and some classical indexes of readability are much more effective in the monolingual and the multilingual experiments than what Vajjala and Rama concluded and are even the best performing features when the cross-lingual task is seen as a regression problem. This study emphasized the importance for reproducibility of setting explicitly the reading order of the instances when using a K-fold CV procedure and, more generally, the need to properly randomize these instances before. It also evaluates a two-step procedure to determine the degree of statistical significance of the differences observed in a K-fold cross-validation schema and argues against the use of a Bonferroni-type correction in this context.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.687}
}

@InProceedings{huber-ltekin:2020:LREC,
  author   = {Huber, Eva and Çöltekin, Çağrı},
  title   = {Reproduction and Replication: A Case Study with Automatic Essay Scoring},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5603--5613},
  abstract = {As in many experimental sciences, reproducibility of experiments has gained ever more attention in the NLP community. This paper presents our reproduction efforts of an earlier study of automatic essay scoring (AES) for determining the proficiency of second language learners in a multilingual setting. We present three sets of experiments with different objectives. First, as prescribed by the LREC 2020 REPROLANG shared task, we rerun the original AES system using the code published by the original authors on the same dataset. Second, we repeat the same experiments on the same data with a different implementation. And third, we test the original system on a different dataset and a different language. Most of our findings are in line with the findings of the original paper. Nevertheless, there are some discrepancies between our results and the results presented in the original paper. We report and discuss these differences in detail. We further go into some points related to confirmation of research findings through reproduction, including the choice of the dataset, reporting and accounting for variability, use of appropriate evaluation metrics, and making code and data available. We also discuss the varying uses and differences between the terms reproduction and replication, and we argue that reproduction, the confirmation of conclusions through independent experiments in varied settings is more valuable than exact replication of the published values.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.688}
}

```
@InProceedings{caines-buttery:2020:LREC,
  author    = {Caines, Andrew  and  Buttery, Paula},
  title     = {REPROLANG 2020: Automatic Proficiency Scoring of
Czech, English, German, Italian, and Spanish Learner Essays},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5614--5623},
  abstract  = {We report on our attempts to reproduce the work
described in Vajjala \& Rama 2018, `Experiments with universal CEFR
classification', as part of REPROLANG 2020: this involves featured-
based and neural approaches to essay scoring in Czech, German and
Italian. Our results are broadly in line with those from the
original paper, with some differences due to the stochastic nature
of machine learning and programming language used. We correct an
error in the reported metrics, introduce new baselines, apply the
experiments to English and Spanish corpora, and generate adversarial
data to test classifier robustness. We conclude that feature-based
approaches perform better than neural network classifiers for text
datasets of this size, though neural network modifications do bring
performance closer to the best feature-based models.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.689}
}

@InProceedings{bunt-EtAl:2020:LREC,
  author    = {Bunt, Harry  and  Petukhova, Volha  and  Gilmartin,
Emer  and  Pelachaud, Catherine  and  Fang, Alex  and  Keizer, Simon
and  Prévot, Laurent},
  title     = {The ISO Standard for Dialogue Act Annotation, Second
Edition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {549--558},
  abstract  = {ISO standard 24617-2 for dialogue act annotation,
established in 2012, has in the past few years been used both in
corpus annotation and in the design of components for spoken and
multimodal dialogue systems. This has brought some inaccuracies and
undesirbale limitations of the standard to light, which are
addressed in a proposed second edition. This second edition allows a
more accurate annotation of dependence relations and rhetorical
relations in dialogue. Following the ISO 24617-4 principles of
semantic annotation, and borrowing ideas from EmotionML, a triple-
layered plug-in mechanism is introduced which allows dialogue act
descriptions to be enriched with information about their semantic
content, about accompanying emotions, and other information, and
allows the annotation scheme to be customised by adding application-
specific dialogue act types.},
```

  url   = {https://www.aclweb.org/anthology/2020.lrec-1.69}
}

@InProceedings{arhiliuc-mitrovi-granitzer:2020:LREC,
  author   = {Arhiliuc, Cristina and Mitrović, Jelena and
Granitzer, Michael},
  title   = {Language Proficiency Scoring},
  booktitle   = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5624--5630},
  abstract = {The Common European Framework of Reference (CEFR)
provides generic guidelines for the evaluation of language
proficiency. Nevertheless, for automated proficiency classification
systems, different approaches for different languages are proposed.
Our paper evaluates and extends the results of an approach to
Automatic Essay Scoring proposed as a part of the REPROLANG 2020
challenge. We provide a comparison between our results and the ones
from the published paper and we include a new corpus for the English
language for further experiments. Our results are lower than the
expected ones when using the same approach and the system does not
scale well with the added English corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.690}
}

@InProceedings{ballier-EtAl:2020:LREC,
  author   = {Ballier, Nicolas and Amari, Nabil and Merat,
Laure and Yunès, Jean-Baptiste},
  title   = {The Learnability of the Annotated Input in NMT
Replicating (Vanmassenhove and Way, 2018) with OpenNMT},
  booktitle   = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5631--5640},
  abstract = {In this paper, we reproduce some of the experiments
related to neural network training for Machine Translation as
reported in (Vanmassenhove and Way, 2018). They annotated a sample
from the EN-FR and EN-DE Europarl aligned corpora with syntactic and
semantic annotations to train neural networks with the Nematus
Neural Machine Translation (NMT) toolkit. Following the original
publication, we obtained lower BLEU scores than the authors of the
original paper, but on a more limited set of annotations. In the
second half of the paper, we try to analyze the difference in the
results obtained and suggest some methods to improve the results. We
discuss the Byte Pair Encoding (BPE) used in the pre-processing
phase and suggest feature ablation in relation to the granularity of
syntactic and semantic annotations. The learnability of the
annotated input is discussed in relation to existing resources for

the target languages. We also discuss the feature representation likely to have been adopted for combining features.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.691}
}

@InProceedings{portisch-hladik-paulheim:2020:LREC,
  author   = {Portisch, Jan and Hladik, Michael and Paulheim, Heiko},
  title   = {KGvec2go – Knowledge Graph Embeddings as a Service},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5641--5647},
  abstract = {In this paper, we present KGvec2go, a Web API for accessing and consuming graph embeddings in a light-weight fashion in downstream applications. Currently, we serve pre-trained embeddings for four knowledge graphs. We introduce the service and its usage, and we show further that the trained models have semantic value by evaluating them on multiple semantic benchmarks. The evaluation also reveals that the combination of multiple models can lead to a better outcome than the best individual model.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.692}
}

@InProceedings{bento-zouaq-gagnon:2020:LREC,
  author   = {Bento, Alexandre and Zouaq, Amal and Gagnon, Michel},
  title   = {Ontology Matching Using Convolutional Neural Networks},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5648--5653},
  abstract = {In order to achieve interoperability of information in the context of the Semantic Web, it is necessary to find effective ways to align different ontologies. As the number of ontologies grows for a given domain, and as overlap between ontologies grows proportionally, it is becoming more and more crucial to develop accurate and reliable techniques to perform this task automatically. While traditional approaches to address this challenge are based on string metrics and structure analysis, in this paper we present a methodology to align ontologies automatically using machine learning techniques. Specifically, we use convolutional neural networks to perform string matching between class labels using character embeddings. We also rely on the set of superclasses to perform the best alignment. Our results show that we obtain state-of-the-art performance on ontologies from the Ontology Alignment Evaluation Initiative (OAEI). Our model also maintains

good performance when tested on a different domain, which could lead to potential cross-domain applications.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.693}
}

@InProceedings{martnchozas-ahmadi-montielponsoda:2020:LREC,
  author   = {Martín-Chozas, Patricia and Ahmadi, Sina and Montiel-Ponsoda, Elena},
  title   = {Defying Wikidata: Validation of Terminological Relations in the Web of Data},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5654--5659},
  abstract = {In this paper we present an approach to validate terminological data retrieved from open encyclopaedic knowledge bases. This need arises from the enrichment of automatically extracted terms with information from existing resources in theLinguistic Linked Open Data cloud. Specifically, the resource employed for this enrichment is WIKIDATA, since it is one of the biggest knowledge bases freely available within the Semantic Web. During the experiment, we noticed that certain RDF properties in the Knowledge Base did not contain the data they are intended to represent, but a different type of information. In this paper we propose an approach to validate the retrieved data based on four axioms that rely on two linguistic theories: the x-bar theory and the multidimensional theory of terminology.The validation process is supported by a second knowledge base specialised in linguistic data; in this case, CONCEPTNET. In our experiment, we validate terms from the legal domain in four languages: Dutch, English, German and Spanish. The final aim is to generate a set of sound and reliable terminological resources in RDF to contribute to the population of the Linguistic Linked Open Data cloud.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.694}
}

@InProceedings{declerck-EtAl:2020:LREC,
  author   = {Declerck, Thierry and McCrae, John Philip and Hartung, Matthias and Gracia, Jorge and Chiarcos, Christian and Montiel-Ponsoda, Elena and Cimiano, Philipp and Revenko, Artem and Saurí, Roser and Lee, Deirdre and Racioppa, Stefania and Abdul Nasir, Jamal and Orlikowsk, Matthias and Lanau-Coronas, Marta and Fäth, Christian and Rico, Mariano and Elahi, Mohammad Fazleh and Khvalchik, Maria and Gonzalez, Meritxell and Cooney, Katharine},
  title   = {Recent Developments for the Linguistic Linked Open Data Infrastructure},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5660--5667},
  abstract  = {In this paper we describe the contributions made by
the European H2020 project “Prêt-à-LLOD” (‘Ready-to-use Multilingual
Linked Language Data for Knowledge Services across Sectors’) to the
further development of the Linguistic Linked Open Data (LLOD)
infrastructure. Prêt-à-LLOD aims to develop a new methodology for
building data value chains applicable to a wide range of sectors and
applications and based around language resources and language
technologies that can be integrated by means of semantic
technologies. We describe the methods implemented for increasing the
number of language data sets in the LLOD. We also present the
approach for ensuring interoperability and for porting LLOD data
sets and services to other infrastructures, as well as the
contribution of the projects to existing standards.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.695}
}

@InProceedings{chiarcos-fth-abromeit:2020:LREC,
  author    = {Chiarcos, Christian  and  Fäth, Christian  and
Abromeit, Frank},
  title     = {Annotation Interoperability for the Post-ISOCat Era},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5668--5677},
  abstract  = {With this paper, we provide an overview over ISOCat
successor solutions and annotation standardization efforts since
2010, and we describe the low-cost harmonization of post-ISOCat
vocabularies by means of modular, linked ontologies: The CLARIN
Concept Registry, LexInfo, Universal Parts of Speech, Universal
Dependencies and UniMorph are linked with the Ontologies of
Linguistic Annotation and through it with ISOCat, the GOLD ontology,
the Typological Database Systems ontology and a large number of
annotation schemes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.696}
}

@InProceedings{alexmathews-strube:2020:LREC,
  author    = {Alex Mathews, Kevin  and  Strube, Michael},
  title     = {A Large Harvested Corpus of Location Metonymy},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5678--5687},
  abstract  = {Metonymy is a figure of speech in which an entity is
referred to by another related entity. The existing datasets of
```

metonymy are either too small in size or lack sufficient coverage. We propose a new, labelled, high-quality corpus of location metonymy called WiMCor, which is large in size and has high coverage. The corpus is harvested semi-automatically from English Wikipedia. We use different labels of varying granularity to annotate the corpus. The corpus can directly be used for training and evaluating automatic metonymy resolution systems. We construct benchmarks for metonymy resolution, and evaluate baseline methods using the new corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.697}
}

@InProceedings{robaldo-bartolini-lenzini:2020:LREC,
  author   = {Robaldo, Livio and Bartolini, Cesare and Lenzini, Gabriele},
  title   = {The DAPRECO Knowledge Base: Representing the GDPR in LegalRuleML},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5688--5697},
  abstract = {The DAPRECO knowledge base (D-KB) is a repository of rules written in LegalRuleML, an XML formalism designed to represent the logical content of legal documents. The rules represent the provisions of the General Data Protection Regulation (GDPR). The D-KB builds upon the Privacy Ontology (PrOnto) (Palmirani et al., 2018), which provides a model for the legal concepts involved in the GDPR, by adding a further layer of constraints in the form of if-then rules, referring either to standard first order logic implications or to deontic statements. If-then rules are formalized in reified I/O logic (Robaldo and Sun, 2017) and then codified in (LegalRuleML, 2019). To date, the D-KB is the biggest knowledge base in LegalRuleML freely available online at (Robaldo et al., 2019).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.698}
}

@InProceedings{white-EtAl:2020:LREC,
  author   = {White, Aaron Steven and Stengel-Eskin, Elias and Vashishtha, Siddharth and Govindarajan, Venkata Subrahmanyan and Reisinger, Dee Ann and Vieira, Tim and Sakaguchi, Keisuke and Zhang, Sheng and Ferraro, Francis and Rudinger, Rachel and Rawlins, Kyle and Van Durme, Benjamin},
  title   = {The Universal Decompositional Semantics Dataset and Decomp Toolkit},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5698--5707},

```
  abstract  = {We present the Universal Decompositional Semantics
(UDS) dataset (v1.0), which is bundled with the Decomp toolkit
(v0.1). UDS1.0 unifies five high-quality, decompositional semantics-
aligned annotation sets within a single semantic graph
specification---with graph structures defined by the predicative
patterns produced by the PredPatt tool and real-valued node and edge
attributes constructed using sophisticated normalization procedures.
The Decomp toolkit provides a suite of Python 3 tools for querying
UDS graphs using SPARQL. Both UDS1.0 and Decomp0.1 are publicly
available at http://decomp.io.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.699}
}

@InProceedings{pagel-reiter:2020:LREC,
  author    = {Pagel, Janis  and  Reiter, Nils},
  title     = {GerDraCor-Coref: A Coreference Corpus for Dramatic
Texts in German},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {55--64},
  abstract  = {Dramatic texts are a highly structured literary text
type. Their quantitative analysis so far has relied on analysing
structural properties (e.g., in the form of networks). Resolving
coreferences is crucial for an analysis of the content of the
character speech, but developing automatic coreference resolution
(CR) systems depends on the existence of annotated corpora. In this
paper, we present an annotated corpus of German dramatic texts, a
preliminary analysis of the corpus as well as some baseline
experiments on automatic CR. The analysis shows that with respect to
the reference structure, dramatic texts are very different from news
texts, but more similar to other dialogical text types such as
interviews. Baseline experiments show a performance of 28.8 CoNLL
score achieved by the rule-based CR system CorZu. In the future, we
plan to integrate the (partial) information given in the dramatis
personae into the CR model.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.7}
}

@InProceedings{jokinen:2020:LREC,
  author    = {Jokinen, Kristiina},
  title     = {The AICO Multimodal Corpus – Data Collection and
Preliminary Analyses},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {559--564},
  abstract  = {This paper describes data collection and the first
```

```
explorative research on the AICO Multimodal Corpus. The corpus
contains eye-gaze, Kinect, and video recordings of human-robot and
human-human interactions, and was collected to study cooperation,
engagement and attention of human participants in task-based as well
as in chatty type interactive situations. In particular, the goal
was to enable comparison between human-human and human-robot
interactions, besides studying multimodal behaviour and attention in
the different dialogue activities. The robot partner was a humanoid
Nao robot, and it was expected that its agent-like behaviour would
render humanrobot interactions similar to human-human interaction
but also high-light important differences due to the robot's limited
conversational capabilities. The paper reports on the preliminary
studies on the corpus, concerning the participants' eye-gaze and
gesturing behaviours,which were chosen as objective measures to
study differences in their multimodal behaviour patterns with a
human and a robot partner.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.70}
}

@InProceedings{chersoni-EtAl:2020:LREC,
  author    = {Chersoni, Emmanuele  and  Pannitto, Ludovica  and
Santus, Enrico  and  Lenci, Alessandro  and  Huang, Chu-Ren},
  title     = {Are Word Embeddings Really a Bad Fit for the
Estimation of Thematic Fit?},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5708--5713},
  abstract  = {While neural embeddings represent a popular choice
for word representation in a wide variety of NLP tasks, their usage
for thematic fit modeling has been limited, as they have been
reported to lag behind syntax-based count models. In this paper, we
propose a complete evaluation of count models and word embeddings on
thematic fit estimation, by taking into account a larger number of
parameters and verb roles and introducing also dependency-based
embeddings in the comparison. Our results show a complex scenario,
where a determinant factor for the performance seems to be the
availability to the model of reliable syntactic information for
building the distributional representations of the roles.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.700}
}

@InProceedings{xiang-EtAl:2020:LREC2,
  author    = {Xiang, Rong  and  Gao, Xuefeng  and  Long, Yunfei
and  Li, Anran  and  Chersoni, Emmanuele  and  Lu, Qin  and  Huang,
Chu-Ren},
  title     = {Ciron: a New Benchmark Dataset for Chinese Irony
Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5714--5720},
  abstract  = {Automatic Chinese irony detection is a challenging
task, and it has a strong impact on linguistic research. However,
Chinese irony detection often lacks labeled benchmark datasets. In
this paper, we introduce Ciron, the first Chinese benchmark dataset
available for irony detection for machine learning models. Ciron
includes more than 8.7K posts, collected from Weibo, a micro
blogging platform. Most importantly, Ciron is collected with no pre-
conditions to ensure a much wider coverage. Evaluation on seven
different machine learning classifiers proves the usefulness of
Ciron as an important resource for Chinese irony detection.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.701}
}

@InProceedings{anthonio-bhat-roth:2020:LREC,
  author    = {Anthonio, Talita  and  Bhat, Irshad  and  Roth,
Michael},
  title     = {wikiHowToImprove: A Resource and Analyses on Edits in
Instructional Texts},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5721--5729},
  abstract  = {Instructional texts, such as articles in wikiHow,
describe the actions necessary to accomplish a certain goal. In
wikiHow and other resources, such instructions are subject to
revision edits on a regular basis. Do these edits improve
instructions only in terms of style and correctness, or do they
provide clarifications necessary to follow the instructions and to
accomplish the goal? We describe a resource and first studies
towards answering this question. Specifically, we create
wikiHowToImprove, a collection of revision histories for about 2.7
million sentences from about 246\,000 wikiHow articles. We describe
human annotation studies on categorizing a subset of sentence-level
edits and provide baseline models for the task of automatically
distinguishing ``older'' from ``newer'' revisions of a sentence.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.702}
}

@InProceedings{king-morante:2020:LREC,
  author    = {King, Liza  and  Morante, Roser},
  title     = {Must Children be Vaccinated or not? Annotating Modal
Verbs in the Vaccination Debate},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {5730--5738},
  abstract  = {In this paper we analyze the use of modal verbs in a
corpus of texts related to the vaccination debate. Broadly speaking,
the vaccination debate centers around whether vaccination is safe,
and whether it is morally acceptable to enforce mandatory
vaccination. In order to successfully intervene and curb the spread
of preventable diseases due to low vaccination rates, health
practitioners need to be adequately informed on public perception of
the safety and necessity of vaccines. Public perception can relate
to the strength of conviction that an individual may have towards a
proposition (e.g. `one must vaccinate' versus `one should
vaccinate'), as well as qualify the type of proposition, be it
related to morality (`government should not interfere in my personal
choice') or related to possibility (`too many vaccines at once could
hurt my child'). Text mining and analysis of modal auxiliaries are
economically viable means of gaining insights into these
perspectives, particularly on a large scale due to the widespread
use of social media and blogs as vehicles of communication.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.703}
}

@InProceedings{khandelwal-sawant:2020:LREC,
  author    = {Khandelwal, Aditya  and  Sawant, Suraj},
  title     = {NegBERT: A Transfer Learning Approach for Negation
Detection and Scope Resolution},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5739--5748},
  abstract  = {Negation is an important characteristic of language,
and a major component of information extraction from text. This
subtask is of considerable importance to the biomedical domain. Over
the years, multiple approaches have been explored to address this
problem: Rule-based systems, Machine Learning classifiers,
Conditional Random Field models, CNNs and more recently BiLSTMs. In
this paper, we look at applying Transfer Learning to this problem.
First, we extensively review previous literature addressing Negation
Detection and Scope Resolution across the 3 datasets that have
gained popularity over the years: the BioScope Corpus, the Sherlock
dataset, and the SFU Review Corpus. We then explore the decision
choices involved with using BERT, a popular transfer learning model,
for this task, and report state-of-the-art results for scope
resolution across all 3 datasets. Our model, referred to as NegBERT,
achieves a token level F1 score on scope resolution of 92.36 on the
Sherlock dataset, 95.68 on the BioScope Abstracts subcorpus, 91.24
on the BioScope Full Papers subcorpus, 90.95 on the SFU Review
Corpus, outperforming the previous state-of-the-art systems by a
significant margin. We also analyze the model’s generalizability to
datasets on which it is not trained.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.704}
```

}

@InProceedings{majewska-EtAl:2020:LREC,
  author   = {Majewska, Olga and McCarthy, Diana and van den Bosch, Jasper and Kriegeskorte, Nikolaus and Vulić, Ivan and Korhonen, Anna},
  title   = {Spatial Multi-Arrangement for Clustering and Multi-way Similarity Dataset Construction},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5749--5758},
  abstract = {We present a novel methodology for fast bottom-up creation of large-scale semantic similarity resources to support development and evaluation of NLP systems. Our work targets verb similarity, but the methodology is equally applicable to other parts of speech. Our approach circumvents the bottleneck of slow and expensive manual development of lexical resources by leveraging semantic intuitions of native speakers and adapting a spatial multi-arrangement approach from cognitive neuroscience, used before only with visual stimuli, to lexical stimuli. Our approach critically obtains judgments of word similarity in the context of a set of related words, rather than of word pairs in isolation. We also handle lexical ambiguity as a natural consequence of a two-phase process where verbs are placed in broad semantic classes prior to the fine-grained spatial similarity judgments. Our proposed design produces a large-scale verb resource comprising 17 relatedness-based classes and a verb similarity dataset containing similarity scores for 29,721 unique verb pairs and 825 target verbs, which we release with this paper.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.705}
}

@InProceedings{pasini-camachocollados:2020:LREC,
  author   = {Pasini, Tommaso and Camacho-Collados, Jose},
  title   = {A Short Survey on Sense-Annotated Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5759--5765},
  abstract = {Large sense-annotated datasets are increasingly necessary for training deep supervised systems in Word Sense Disambiguation. However, gathering high-quality sense-annotated data for as many instances as possible is a laborious and expensive task. This has led to the proliferation of automatic and semi-automatic methods for overcoming the so-called knowledge-acquisition bottleneck. In this short survey we present an overview of sense-annotated corpora, annotated either manually- or

(semi)automatically, that are currently available for different languages and featuring distinct lexical resources as inventory of senses, i.e. WordNet, Wikipedia, BabelNet. Furthermore, we provide the reader with general statistics of each dataset and an analysis of their specific features.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.706}
}

@InProceedings{jana-varimalla-goyal:2020:LREC,
  author   = {Jana, Abhik and Varimalla, Nikhil Reddy and Goyal, Pawan},
  title   = {Using Distributional Thesaurus Embedding for Co-hyponymy Detection},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5766--5771},
  abstract = {Discriminating lexical relations among distributionally similar words has always been a challenge for natural language processing (NLP) community. In this paper, we investigate whether the network embedding of distributional thesaurus can be effectively utilized to detect co-hyponymy relations. By extensive experiments over three benchmark datasets, we show that the vector representation obtained by applying node2vec on distributional thesaurus outperforms the state-of-the-art models for binary classification of co-hyponymy vs. hypernymy, as well as co-hyponymy vs. meronymy, by huge margins.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.707}
}

@InProceedings{limalopez-EtAl:2020:LREC1,
  author   = {Lima Lopez, Salvador and Perez, Naiara and Cuadros, Montse and Rigau, German},
  title   = {NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5772--5781},
  abstract = {This paper introduces the first version of the NUBes corpus (Negation and Uncertainty annotations in Biomedical texts in Spanish). The corpus is part of an on-going research and currently consists of 29,682 sentences obtained from anonymised health records annotated with negation and uncertainty. The article includes an exhaustive comparison with similar corpora in Spanish, and presents the main annotation and design decisions. Additionally, we perform preliminary experiments using deep learning algorithms to validate the annotated dataset. As far as we know, NUBes is the largest

```
available corpora for negation in Spanish and the first that also
incorporates the annotation of speculation cues, scopes, and
events.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.708}
}

@InProceedings{kovatchev-EtAl:2020:LREC,
  author    = {Kovatchev, Venelin  and  Gold, Darina  and  Marti, M.
Antonia  and  Salamo, Maria  and  Zesch, Torsten},
  title     = {Decomposing and Comparing Meaning Relations:
Paraphrasing, Textual Entailment, Contradiction, and Specificity},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5782--5791},
  abstract  = {In this paper, we present a methodology for
decomposing and comparing multiple meaning relations (paraphrasing,
textual entailment, contradiction, and specificity). The methodology
includes SHARel - a new typology that consists of 26 linguistic and
8 reason-based categories. We use the typology to annotate a corpus
of 520 sentence pairs in English and we demonstrate that unlike
previous typologies, SHARel can be applied to all relations of
interest with a high inter-annotator agreement. We analyze and
compare the frequency and distribution of the linguistic and reason-
based phenomena involved in paraphrasing, textual entailment,
contradiction, and specificity. This comparison allows for a much
more in-depth analysis of the workings of the individual relations
and the way they interact and compare with each other. We release
all resources (typology, annotation guidelines, and annotated
corpus) to the community.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.709}
}

@InProceedings{galetzka-eneh-schlangen:2020:LREC,
  author    = {Galetzka, Fabian  and  Eneh, Chukwuemeka Uchenna  and
Schlangen, David},
  title     = {A Corpus of Controlled Opinionated and Knowledgeable
Movie Discussions for Training Neural Conversation Models},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {565--573},
  abstract  = {Fully data driven Chatbots for non-goal oriented
dialogues are known to suffer from inconsistent behaviour across
their turns, stemming from a general difficulty in controlling
parameters like their assumed background personality and knowledge
of facts. One reason for this is the relative lack of labeled data
from which personality consistency and fact usage could be learned
```

together with dialogue behaviour. To address this, we introduce a new labeled dialogue dataset in the domain of movie discussions, where every dialogue is based on pre-specified facts and opinions. We thoroughly validate the collected dialogue for adherence of the participants to their given fact and opinion profile, and find that the general quality in this respect is high. This process also gives us an additional layer of annotation that is potentially useful for training models. We introduce as a baseline an end-to-end trained self-attention decoder model trained on this data and show that it is able to generate opinionated responses that are judged to be natural and knowledgeable and show attentiveness.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.71}
}

@InProceedings{silberer-zarrie-boleda:2020:LREC,
  author   = {Silberer, Carina and Zarrieß, Sina and Boleda, Gemma},
  title   = {Object Naming in Language and Vision: A Survey and a New Dataset},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5792--5801},
  abstract = {People choose particular names for objects, such as dog or puppy for a given dog. Object naming has been studied in Psycholinguistics, but has received relatively little attention in Computational Linguistics. We review resources from Language and Vision that could be used to study object naming on a large scale, discuss their shortcomings, and create a new dataset that affords more opportunities for analysis and modeling. Our dataset, ManyNames, provides 36 name annotations for each of 25K objects in images selected from VisualGenome. We highlight the challenges involved and provide a preliminary analysis of the ManyNames data, showing that there is a high level of agreement in naming, on average. At the same time, the average number of name types associated with an object is much higher in our dataset than in existing corpora for Language and Vision, such that ManyNames provides a rich resource for studying phenomena like hierarchical variation (chihuahua vs. dog), which has been discussed at length in the theoretical literature, and other less well studied phenomena like cross-classification (cake vs. dessert).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.710}
}

@InProceedings{yen-EtAl:2020:LREC,
  author   = {Yen, Ting-Yu and Lee, Yang-Yin and Shiue, Yow-Ting and Huang, Hen-Hsen and Chen, Hsin-Hsi},
  title   = {MSD-1030: A Well-built Multi-Sense Evaluation Dataset for Sense Representation Models},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5802--5809},
  abstract  = {Sense embedding models handle polysemy by giving each
distinct meaning of a word form a separate representation. They are
considered improvements over word models, and their effectiveness is
usually judged with benchmarks such as semantic similarity datasets.
However, most of these datasets are not designed for evaluating
sense embeddings. In this research, we show that there are at least
six concerns about evaluating sense embeddings with existing
benchmark datasets, including the large proportions of single-sense
words and the unexpected inferior performance of several multi-sense
models to their single-sense counterparts. These observations call
into serious question whether evaluations based on these datasets
can reflect the sense model's ability to capture different meanings.
To address the issues, we propose the Multi-Sense Dataset
(MSD-1030), which contains a high ratio of multi-sense word pairs. A
series of analyses and experiments show that MSD-1030 serves as a
more reliable benchmark for sense embeddings. The dataset is
available at http://nlg.csie.ntu.edu.tw/nlpresource/MSD-1030/.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.711}
}

@InProceedings{zayed-mccrae-buitelaar:2020:LREC,
  author    = {Zayed, Omnia  and  McCrae, John Philip  and
Buitelaar, Paul},
  title     = {Figure Me Out: A Gold Standard Dataset for Metaphor
Interpretation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5810--5819},
  abstract  = {Metaphor comprehension and understanding is a complex
cognitive task that requires interpreting metaphors by grasping the
interaction between the meaning of their target and source concepts.
This is very challenging for humans, let alone computers. Thus,
automatic metaphor interpretation is understudied in part due to the
lack of publicly available datasets. The creation and manual
annotation of such datasets is a demanding task which requires huge
cognitive effort and time. Moreover, there will always be a question
of accuracy and consistency of the annotated data due to the
subjective nature of the problem. This work addresses these issues
by presenting an annotation scheme to interpret verb-noun metaphoric
expressions in text. The proposed approach is designed with the goal
of reducing the workload on annotators and maintain consistency. Our
methodology employs an automatic retrieval approach which utilises
external lexical resources, word embeddings and semantic similarity
to generate possible interpretations of identified metaphors in
order to enable quick and accurate annotation. We validate our
```

proposed approach by annotating around 1,500 metaphors in tweets which were annotated by six native English speakers. As a result of this work, we publish as linked data the first gold standard dataset for metaphor interpretation which will facilitate research in this area.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.712}
}

@InProceedings{tanguy-brunet-ferret:2020:LREC,
  author   = {Tanguy, Ludovic and Brunet, Pauline and Ferret, Olivier},
  title   = {Extrinsic Evaluation of French Dependency Parsers on a Specialized Corpus: Comparison of Distributional Thesauri},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5820--5828},
  abstract = {We present a study in which we compare 11 different French dependency parsers on a specialized corpus (consisting of research articles on NLP from the proceedings of the TALN conference). Due to the lack of a suitable gold standard, we use each of the parsers' output to generate distributional thesauri using a frequency-based method. We compare these 11 thesauri to assess the impact of choosing a parser over another. We show that, without any reference data, we can still identify relevant subsets among the different parsers. We also show that the similarity we identify between parsers is confirmed on a restricted distributional benchmark.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.713}
}

@InProceedings{yu-EtAl:2020:LREC,
  author   = {Yu, Xiaojing and Chen, Tianlong and Yu, Zhengjie and Li, Huiyu and Yang, Yang and Jiang, Xiaoqian and Jiang, Anxiao},
  title   = {Dataset and Enhanced Model for Eligibility Criteria-to-SQL Semantic Parsing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5829--5837},
  abstract = {Clinical trials often require that patients meet eligibility criteria (e.g., have specific conditions) to ensure the safety and the effectiveness of studies. However, retrieving eligible patients for a trial from the electronic health record (EHR) database remains a challenging task for clinicians since it requires not only medical knowledge about eligibility criteria, but also an adequate understanding of structured query language (SQL).

```
In this paper, we introduce a new dataset that includes the first-of-its-kind eligibility-criteria corpus and the corresponding queries for criteria-to-sql (Criteria2SQL), a task translating the eligibility criteria to executable SQL queries. Compared to existing datasets, the queries in the dataset here are derived from the eligibility criteria of clinical trials and include {\it Order-sensitive, Counting-based, and Boolean-type} cases which are not seen before. In addition to the dataset, we propose a novel neural semantic parser as a strong baseline model. Extensive experiments show that the proposed parser outperforms existing state-of-the-art general-purpose text-to-sql models while highlighting the challenges presented by the new dataset. The uniqueness and the diversity of the dataset leave a lot of research opportunities for future improvement.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.714}
}

@InProceedings{roussinov-sharoff-puchnina:2020:LREC,
  author    = {Roussinov, Dmitri  and  Sharoff, Serge  and  Puchnina, Nadezhda},
  title     = {Recognizing Semantic Relations by Combining Transformers and Fully Connected Models},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5838--5845},
  abstract  = {Automatically recognizing an existing semantic relation (e.g. “is a”, “part of”, “property of”, “opposite of” etc.) between two words (phrases, concepts, etc.) is an important task affecting many NLP applications and has been subject of extensive experimentation and modeling. Current approaches to automatically telling if a relation exists between two given concepts X and Y can be grouped into two types: 1) those modeling word-paths connecting X and Y in text and 2) those modeling distributional properties of X and Y separately, not necessary in the proximity to each other. Here, we investigate how both types can be improved and combined. We suggest a distributional approach that is based on an attention-based transformer. We have also developed a novel word path model that combines useful properties of a convolutional network with a fully connected language model. While our transformer-based approach works better, both our models significantly outperform the state-of-the-art within their classes of approaches. We also demonstrate that combining the two approaches results in additional gains since they use somewhat different data sources.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.715}
}

@InProceedings{hasegawa-kobayashi-hayashi:2020:LREC,
  author    = {Hasegawa, Mika  and  Kobayashi, Tetsunori  and  Hayashi, Yoshihiko},
  title     = {Word Attribute Prediction Enhanced by Lexical
```

Entailment Tasks},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5846--5854},
  abstract = {Human semantic knowledge about concepts acquired through perceptual inputs and daily experiences can be expressed as a bundle of attributes. Unlike the conventional distributed word representations that are purely induced from a text corpus, a semantic attribute is associated with a designated dimension in attribute-based vector representations. Thus, semantic attribute vectors can effectively capture the commonalities and differences among concepts. However, as semantic attributes have been generally created by psychological experimental settings involving human annotators, an automatic method to create or extend such resources is highly demanded in terms of language resource development and maintenance. This study proposes a two-stage neural network architecture, Word2Attr, in which initially acquired attribute representations are then fine-tuned by employing supervised lexical entailment tasks. The quantitative empirical results demonstrated that the fine-tuning was indeed effective in improving the performances of semantic/visual similarity/relatedness evaluation tasks. Although the qualitative analysis confirmed that the proposed method could often discover valid but not-yet human-annotated attributes, they also exposed future issues to be worked: we should refine the inventory of semantic attributes that currently relies on an existing dataset.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.716}
}

@InProceedings{dan-EtAl:2020:LREC,
  author   = {Dan, Soham and Kordjamshidi, Parisa and Bonn, Julia and Bhatia, Archna and Cai, Zheng and Palmer, Martha and Roth, Dan},
  title   = {From Spatial Relations to Spatial Configurations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5855--5864},
  abstract = {Spatial Reasoning from language is essential for natural language understanding. Supporting it requires a representation scheme that can capture spatial phenomena encountered in language as well as in images and videos.Existing spatial representations are not sufficient for describing spatial configurations used in complex tasks. This paper extends the capabilities of existing spatial representation languages and increases coverage of the semantic aspects that are needed to ground spatial meaning of natural language text in the world. Our spatial

relation language is able to represent a large, comprehensive set of spatial concepts crucial for reasoning and is designed to support composition of static and dynamic spatial configurations. We integrate this language with the Abstract Meaning Representation (AMR) annotation schema and present a corpus annotated by this extended AMR. To exhibit the applicability of our representation scheme, we annotate text taken from diverse datasets and show how we extend the capabilities of existing spatial representation languages with fine-grained decomposition of semantics and blend it seamlessly with AMRs of sentences and discourse representations as a whole.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.717}
}

@InProceedings{sucameli-lenci:2020:LREC,
  author   = {Sucameli, Irene and Lenci, Alessandro},
  title   = {Representing Verbs with Visual Argument Vectors},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5865--5870},
  abstract = {Is it possible to use images to model verb semantic similarities? Starting from this core question, we developed two textual distributional semantic models and a visual one. We found particularly interesting and challenging to investigate this Part of Speech since verbs are not often analysed in researches focused on multimodal distributional semantics. After the creation of the visual and textual distributional space, the three models were evaluated in relation to SimLex-999, a gold standard resource. Through this evaluation, we demonstrate that, using visual distributional models, it is possible to extract meaningful information and to effectively capture the semantic similarity between verbs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.718}
}

@InProceedings{mykowiecka-marciniak:2020:LREC,
  author   = {Mykowiecka, Agnieszka and Marciniak, Malgorzata},
  title   = {Are White Ravens Ever White? - Non-Literal Adjective-Noun Phrases in Polish},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5871--5877},
  abstract = {In the paper we describe two resources of Polish data focused on literal and metaphorical meanings of adjective-noun phrases. The first one is FigAN and consists of isolated phrases which are divided into three types: phrases with only literal meaning, with only metaphorical meaning, and phrases which can be

interpreted as literal or metaphorical ones depending on a context of use. The second data is the FigSen corpus which consists of 1833 short fragments of texts containing at least one phrase from the FigAN data which may have both meanings. The corpus is annotated in two ways. One approach concerns annotation of all adjective-noun phrases. In the second approach, literal or metaphorical senses are assigned to all adjectives and nouns in the data. The paper addresses statistics of data and compares two types of annotation. The corpora were used in experiments of automatic recognition of Polish non-literal adjective noun phrases.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.719}
}

@InProceedings{laleye-EtAl:2020:LREC,
  author   = {Laleye, Fréjus A. A. and de Chalendar, Gaël and Blanié, Antonia and Brouquet, Antoine and Behnamou, Dan},
  title   = {A French Medical Conversations Corpus Annotated for a Virtual Patient Dialogue System},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {574--580},
  abstract = {Data-driven approaches for creating virtual patient dialogue systems require the availability of large data specific to the language,domain and clinical cases studied. Based on the lack of dialogue corpora in French for medical education, we propose an annotatedcorpus of dialogues including medical consultation interactions between doctor and patient. In this work, we detail the building processof the proposed dialogue corpus, describe the annotation guidelines and also present the statistics of its contents. We then conducted aquestion categorization task to evaluate the benefits of the proposed corpus that is made publicly available.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.72}
}

@InProceedings{armendariz-EtAl:2020:LREC,
  author   = {Armendariz, Carlos Santos and Purver, Matthew and Ulčar, Matej and Pollak, Senja and Ljubešić, Nikola and Granroth-Wilding, Mark},
  title   = {CoSimLex: A Resource for Evaluating Graded Word Similarity in Context},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5878--5886},
  abstract = {State of the art natural language processing tools are built on context-dependent word embeddings, but no direct method

```
for evaluating these representations currently exists. Standard
tasks and datasets for intrinsic evaluation of embeddings are based
on judgements of similarity, but ignore context; standard tasks for
word sense disambiguation take account of context but do not provide
continuous measures of meaning similarity. This paper describes an
effort to build a new dataset, CoSimLex, intended to fill this gap.
Building on the standard pairwise similarity task of SimLex-999, it
provides context-dependent similarity measures; covers not only
discrete differences in word sense but more subtle, graded changes
in meaning; and covers not only a well-resourced language (English)
but a number of less-resourced languages. We define the task and
evaluation metrics, outline the dataset collection methodology, and
describe the status of the dataset so far.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.720}
}

@InProceedings{amblard-EtAl:2020:LREC,
  author    = {Amblard, Maxime  and  Beysson, Clément  and  de
Groote, Philippe  and  Guillaume, Bruno  and  Pogodalla, Sylvain},
  title     = {A French Version of the FraCaS Test Suite},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5887--5895},
  abstract  = {This paper presents a French version of the FraCaS
test suite. This test suite, originally written in English, contains
problems illustrating semantic inference in natural language. We
describe linguistic choices we had to make when translating the
FraCaS test suite in French, and discuss some of the issues that
were raised by the translation. We also report an experiment we ran
in order to test both the translation and the logical semantics
underlying the problems of the test suite. This provides a way of
checking formal semanticists' hypotheses against actual semantic
capacity of speakers (in the present case, French speakers), and
allow us to compare the results we obtained with the ones of similar
experiments that have been conducted for other languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.721}
}

@InProceedings{yimam-EtAl:2020:LREC,
  author    = {Yimam, Seid Muhie  and  Venkatesh, Gopalakrishnan
and  Lee, John  and  Biemann, Chris},
  title     = {Automatic Compilation of Resources for Academic
Writing and Evaluating with Informal Word Identification and
Paraphrasing System},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {5896--5904},
  abstract  = {We present the first approach to automatically
building resources for academic writing. The aim is to build a
writing aid system that automatically edits a text so that it better
adheres to the academic style of writing. On top of existing
academic resources, such as the Corpus of Contemporary American
English (COCA) academic Word List, the New Academic Word List, and
the Academic Collocation List, we also explore how to dynamically
build such resources that would be used to automatically identify
informal or non-academic words or phrases. The resources are
compiled using different generic approaches that can be extended for
different domains and languages. We describe the evaluation of
resources with a system implementation. The system consists of an
informal word identification (IWI), academic candidate paraphrase
generation, and paraphrase ranking components. To generate
candidates and rank them in context, we have used the PPDB and
WordNet paraphrase resources. We use the Concepts in Context
(CoInCO) "All-Words" lexical substitution dataset both for the
informal word identification and paraphrase generation experiments.
Our informal word identification component achieves an F-1 score of
82\%, significantly outperforming a stratified classifier baseline.
The main contribution of this work is a domain-independent
methodology to build targeted resources for writing aids.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.722}
}

@InProceedings{scarlini-pasini-navigli:2020:LREC,
  author    = {Scarlini, Bianca  and  Pasini, Tommaso  and  Navigli,
Roberto},
  title     = {Sense-Annotated Corpora for Word Sense Disambiguation
in Multiple Languages and Domains},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5905--5911},
  abstract  = {The knowledge acquisition bottleneck problem
dramatically hampers the creation of sense-annotated data for Word
Sense Disambiguation (WSD). Sense-annotated data are scarce for
English and almost absent for other languages. This limits the range
of action of deep-learning approaches, which today are at the base
of any NLP task and are hungry for data. We mitigate this issue and
encourage further research in multilingual WSD by releasing to the
NLP community five large datasets annotated with word-senses in five
different languages, namely, English, French, Italian, German and
Spanish, and 5 distinct datasets in English, each for a different
semantic domain. We show that supervised WSD models trained on our
data attain higher performance than when trained on other
automatically-created corpora. We release all our data containing
more than 15 million annotated instances in 5 different languages at
http://trainomatic.org/onesec.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.723}
```

}

@InProceedings{barque-EtAl:2020:LREC,
  author   = {Barque, Lucie and Haas, Pauline and Huyghe, Richard and Tribout, Delphine and Candito, Marie and Crabbé, Benoit and Segonne, Vincent},
  title   = {FrSemCor: Annotating a French Corpus with Supersenses},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5912--5918},
  abstract = {French, as many languages, lacks semantically annotated corpus data. Our aim is to provide the linguistic and NLP research communities with a gold standard sense-annotated corpus of French, using WordNet Unique Beginners as semantic tags, thus allowing for interoperability. In this paper, we report on the first phase of the project, which focused on the annotation of common nouns. The resulting dataset consists of more than 12,000 French noun occurrences which were annotated in double blind and adjudicated according to a carefully redefined set of supersenses. The resource is released online under a Creative Commons Licence.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.724}
}

@InProceedings{krishnaswamy-pustejovsky:2020:LREC,
  author   = {Krishnaswamy, Nikhil and Pustejovsky, James},
  title   = {A Formal Analysis of Multimodal Referring Strategies Under Common Ground},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5919--5927},
  abstract = {In this paper, we present an analysis of computationally generated mixed-modality definite referring expressions using combinations of gesture and linguistic descriptions. In doing so, we expose some striking formal semantic properties of the interactions between gesture and language, conditioned on the introduction of content into the common ground between the (computational) speaker and (human) viewer, and demonstrate how these formal features can contribute to training better models to predict viewer judgment of referring expressions, and potentially to the generation of more natural and informative referring expressions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.725}
}

@InProceedings{kehat-pustejovsky:2020:LREC,

```
  author    = {Kehat, Gitit  and  Pustejovsky, James},
  title     = {Improving Neural Metaphor Detection with Visual
Datasets},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5928--5933},
  abstract  = {We present new results on Metaphor Detection by using
text from visual datasets. Using a straightforward technique for
sampling text from Vision-Language datasets, we create a data
structure we term a visibility word embedding. We then combine these
embeddings in a relatively simple BiLSTM module augmented with
contextualized word representations (ELMo), and show improvement
over previous state-of-the-art approaches that use more complex
neural network architectures and richer linguistic features, for the
task of verb classification.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.726}
}

@InProceedings{eyal-elhadad:2020:LREC,
  author    = {Eyal, Ben  and  Elhadad, Michael},
  title     = {Building a Hebrew Semantic Role Labeling Lexical
Resource from Parallel Movie Subtitles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5934--5942},
  abstract  = {We present a semantic role labeling resource for
Hebrew built semi-automatically through annotation projection from
English. This corpus is derived from the multilingual OpenSubtitles
dataset and includes short informal sentences, for which reliable
linguistic annotations have been computed. We provide a fully
annotated version of the data including morphological analysis,
dependency syntax and semantic role labeling in both FrameNet and
ProbBank styles. Sentences are aligned between English and Hebrew,
both sides include full annotations and the explicit mapping from
the English arguments to the Hebrew ones. We train a neural SRL
model on this Hebrew resource exploiting the pre-trained
multilingual BERT transformer model, and provide the first available
baseline model for Hebrew SRL as a reference point. The code we
provide is generic and can be adapted to other languages to
bootstrap SRL resources.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.727}
}

@InProceedings{logacheva-EtAl:2020:LREC,
  author    = {Logacheva, Varvara  and  Teslenko, Denis  and
Shelmanov, Artem  and  Remus, Steffen  and  Ustalov, Dmitry  and
```

Kutuzov, Andrey and Artemova, Ekaterina and Biemann, Chris and Ponzetto, Simone Paolo and Panchenko, Alexander},
  title   = {Word Sense Disambiguation for 158 Languages using Word Embeddings Only},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5943--5952},
  abstract = {Disambiguation of word senses in context is easy for humans, but is a major challenge for automatic approaches. Sophisticated supervised and knowledge-based models were developed to solve this task. However, (i) the inherent Zipfian distribution of supervised training instances for a given word and/or (ii) the quality of linguistic knowledge representations motivate the development of completely unsupervised and knowledge-free approaches to word sense disambiguation (WSD). They are particularly useful for under-resourced languages which do not have any resources for building either supervised and/or knowledge-based models. In this paper, we present a method that takes as input a standard pre-trained word embedding model and induces a fully-fledged word sense inventory, which can be used for disambiguation in context. We use this method to induce a collection of sense inventories for 158 languages on the basis of the original pre-trained fastText word embeddings by Grave et al., (2018), enabling WSD in these languages. Models and system are available online.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.728}
}

@InProceedings{sanmartn-trekker-lenaraz:2020:LREC,
  author   = {San Martín, Antonio and Trekker, Catherine and León-Araúz, Pilar},
  title   = {Extraction of Hyponymic Relations in French with Knowledge-Pattern-Based Word Sketches},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5953--5961},
  abstract = {Hyponymy is the cornerstone of taxonomies and concept hierarchies. However, the extraction of hypernym-hyponym pairs from a corpus can be time-consuming, and reconstructing the hierarchical network of a domain is often an extremely complex process. This paper presents the development and evaluation of the French EcoLexicon Semantic Sketch Grammar (ESSG-fr), a French hyponymic sketch grammar for Sketch Engine based on knowledge patterns. It offers a user-friendly way of extracting hyponymic pairs in the form of word sketches in any user-owned corpus. The ESSG-fr contains three times more hyponymic patterns than its English counterpart and has been tested in a multidisciplinary corpus. It is thus expected

```
to be domain-independent. Moreover, the following methodological
innovations have been included in its development: (1) use of
English hyponymic patterns in a parallel corpus to find new French
patterns; (2) automatic inclusion of the results of the Sketch
Engine thesaurus to find new variants of the patterns. As for its
evaluation, the ESSG-fr returns 70\% valid hyperonyms and hyponyms,
measured on 180 extracted pairs of terms in three different
domains.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.729}
}

@InProceedings{wu-EtAl:2020:LREC1,
  author    = {Wu, Chien-Sheng  and  Madotto, Andrea  and  Lin,
Zhaojiang  and  Xu, Peng  and  Fung, Pascale},
  title     = {Getting To Know You: User Attribute Extraction from
Dialogues},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {581--589},
  abstract  = {User attributes provide rich and useful information
for user understanding, yet structured and easy-to-use attributes
are often sparsely populated. In this paper, we leverage dialogues
with conversational agents, which contain strong suggestions of user
information, to automatically extract user attributes. Since no
existing dataset is available for this purpose, we apply distant
supervision to train our proposed two-stage attribute extractor,
which surpasses several retrieval and generation baselines on human
evaluation. Meanwhile, we discuss potential applications (e.g.,
personalized recommendation and dialogue systems) of such extracted
user attributes, and point out current limitations to cast light on
future work.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.73}
}

@InProceedings{strohmaier-EtAl:2020:LREC,
  author    = {Strohmaier, David  and  Gooding, Sian  and
Taslimipoor, Shiva  and  Kochmar, Ekaterina},
  title     = {SeCoDa: Sense Complexity Dataset},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5962--5967},
  abstract  = {The Sense Complexity Dataset (SeCoDa) provides a
corpus that is annotated jointly for complexity and word senses. It
thus provides a valuable resource for both word sense disambiguation
and the task of complex word identification. The intention is that
this dataset will be used to identify complexity at the level of
```

word senses rather than word tokens. For word sense annotation SeCoDa uses a hierarchical scheme that is based on information available in the Cambridge Advanced Learner's Dictionary. This way we can offer more coarse-grained senses than directly available in WordNet.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.730}
}

@InProceedings{rehbein-ruppenhofer:2020:LREC,
  author   = {Rehbein, Ines and Ruppenhofer, Josef},
  title   = {A New Resource for German Causal Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5968--5977},
  abstract = {We present a new resource for German causal language, with annotations in context for verbs, nouns and prepositions. Our dataset includes 4,390 annotated instances for more than 150 different triggers. The annotation scheme distinguishes three different types of causal events (CONSEQUENCE , MOTIVATION, PURPOSE). We also provide annotations for semantic roles, i.e. of the cause and effect for the causal event as well as the actor and affected party, if present. In the paper, we present inter-annotator agreement scores for our dataset and discuss problems for annotating causal language. Finally, we present experiments where we frame causal annotation as a sequence labelling problem and report baseline results for the prediciton of causal arguments and for predicting different types of causation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.731}
}

@InProceedings{choubey-huang:2020:LREC,
  author   = {Choubey, Prafulla Kumar and Huang, Ruihong},
  title   = {One Classifier for All Ambiguous Words: Overcoming Data Sparsity by Utilizing Sense Correlations Across Words},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5978--5985},
  abstract = {Most supervised word sense disambiguation (WSD) systems build word-specific classifiers by leveraging labeled data. However, when using word-specific classifiers, the sparseness of annotations leads to inferior sense disambiguation performance on less frequently seen words. To combat data sparsity, we propose to learn a single model that derives sense representations and meanwhile enforces congruence between a word instance and its right sense by using both sense-annotated data and lexical resources. The model is shared across words that allows utilizing sense

```
correlations across words, and therefore helps to transfer common disambiguation rules from annotation-rich words to annotation-lean words. Empirical evaluation on benchmark datasets shows that the proposed shared model outperforms the equivalent classifier-based models by 1.7\%, 2.5\% and 3.8\% in F1-score when using GloVe, ELMo and BERT word embeddings respectively.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.732}
}

@InProceedings{peng-EtAl:2020:LREC,
  author    = {Peng, Siyao  and  Liu, Yang  and  Zhu, Yilun  and  Blodgett, Austin  and  Zhao, Yushi  and  Schneider, Nathan},
  title     = {A Corpus of Adpositional Supersenses for Mandarin Chinese},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5986--5994},
  abstract  = {Adpositions are frequent markers of semantic relations, but they are highly ambiguous and vary significantly from language to language. Moreover, there is a dearth of annotated corpora for investigating the cross-linguistic variation of adposition semantics, or for building multilingual disambiguation systems. This paper presents a corpus in which all adpositions have been semantically annotated in Mandarin Chinese; to the best of our knowledge, this is the first Chinese corpus to be broadly annotated with adposition semantics. Our approach adapts a framework that defined a general set of supersenses according to ostensibly language-independent semantic criteria, though its development focused primarily on English prepositions (Schneider et al., 2018). We find that the supersense categories are well-suited to Chinese adpositions despite syntactic differences from English. On a Mandarin translation of The Little Prince, we achieve high inter-annotator agreement and analyze semantic correspondences of adposition tokens in bitext.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.733}
}

@InProceedings{moeller-EtAl:2020:LREC,
  author    = {Moeller, Sarah  and  Wagner, Irina  and  Palmer, Martha  and  Conger, Kathryn  and  Myers, Skatje},
  title     = {The Russian PropBank},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {5995--6002},
  abstract  = {This paper presents a proposition bank for Russian (RuPB), a resource for semantic role labeling (SRL). The motivating
```

goal for this resource is to automatically project semantic role labels from English to Russian. This paper describes frame creation strategies, coverage, and the process of sense disambiguation. It discusses language-specific issues that complicated the process of building the PropBank and how these challenges were exploited as language-internal guidance for consistency and coherence.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.734}
}

@InProceedings{jantunen-puupponen-burger:2020:LREC,
  author   = {Jantunen, Tommi and Puupponen, Anna and Burger, Birgitta},
  title   = {What Comes First: Combining Motion Capture and Eye Tracking Data to Study the Order of Articulators in Constructed Action in Sign Language Narratives},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6003--6007},
  abstract = {We use synchronized 120 fps motion capture and 50 fps eye tracking data from two native signers to investigate the temporal order in which the dominant hand, the head, the chest and the eyes start producing overt constructed action from regular narration in seven short Finnish Sign Language stories. From the material, we derive a sample of ten instances of regular narration to overt constructed action transfers in ELAN which we then further process and analyze in Matlab. The results indicate that the temporal order of articulators shows both contextual and individual variation but that there are also repeated patterns which are similar across all the analyzed sequences and signers. Most notably, when the discourse strategy changes from regular narration to overt constructed action, the head and the eyes tend to take the leading role, and the chest and the dominant hand tend to start acting last. Consequences of the findings are discussed.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.735}
}

@InProceedings{naert-larboulette-gibet:2020:LREC,
  author   = {Naert, Lucie and Larboulette, Caroline and Gibet, Sylvie},
  title   = {LSF-ANIMAL: A Motion Capture Corpus in French Sign Language Designed for the Animation of Signing Avatars},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6008--6017},
  abstract = {Signing avatars allow deaf people to access information in their preferred language using an interactive

```
visualization of the sign language spatio-temporal content. However,
avatars are often procedurally animated, resulting in robotic and
unnatural movements, which are therefore rejected by the community
for which they are intended. To overcome this lack of authenticity,
solutions in which the avatar is animated from motion capture data
are promising. Yet, the initial data set drastically limits the
range of signs that the avatar can produce. Therefore, it can be
interesting to enrich the initial corpus with new content by editing
the captured motions. For this purpose, we collected the LSF-ANIMAL
corpus, a French Sign Language (LSF) corpus composed of captured
isolated signs and full sentences that can be used both to study LSF
features and to generate new signs and utterances. This paper
presents the precise definition and content of this corpus,
technical considerations relative to the motion capture process
(including the marker set definition), the post-processing steps
required to obtain data in a standard motion format and the
annotation scheme used to label the data. The quality of the corpus
with respect to intelligibility, accuracy and realism is
perceptually evaluated by 41 participants including native LSF
signers.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.736}
}

@InProceedings{decoster-vanherreweghe-dambre:2020:LREC,
  author    = {De Coster, Mathieu  and  Van Herreweghe, Mieke  and
Dambre, Joni},
  title     = {Sign Language Recognition with Transformer Networks},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6018--6024},
  abstract  = {Sign languages are complex languages. Research into
them is ongoing, supported by large video corpora of which only
small parts are annotated. Sign language recognition can be used to
speed up the annotation process of these corpora, in order to aid
research into sign languages and sign language recognition. Previous
research has approached sign language recognition in various ways,
using feature extraction techniques or end-to-end deep learning. In
this work, we apply a combination of feature extraction using
OpenPose for human keypoint estimation and end-to-end feature
learning with Convolutional Neural Networks. The proven multi-head
attention mechanism used in transformers is applied to recognize
isolated signs in the Flemish Sign Language corpus. Our proposed
method significantly outperforms the previous state of the art of
sign language recognition on the Flemish Sign Language corpus: we
obtain an accuracy of 74.7\% on a vocabulary of 100 classes. Our
results will be implemented as a suggestion system for sign language
corpus annotation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.737}
}
```

@InProceedings{trolvi-delmonte:2020:LREC,
  author   = {Trolvi, Serena and Delmonte, Rodolfo},
  title   = {Annotating a Fable in Italian Sign Language (LIS)},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6025--6034},
  abstract = {This paper introduces work carried out for the automatic generation of a written text in Italian starting from glosses of a fable in Italian Sign Language (LIS). The paper gives a brief overview of sign languages (SLs) and some peculiarities of SL fables such as the use of space, the strategy of Role Shift and classifiers. It also presents the annotation of the fable “The Tortoise and the Hare” - signed in LIS and made available by Alba Cooperativa Sociale -, which was annotated manually by first author for her master’s thesis. The annotation was the starting point of a generation process that allowed us to automatically generate a text in Italian starting from LIS glosses. LIS sentences have been transcribed with Italian words into tables on simultaneous layers, each of which contains specific linguistic or non-linguistic pieces of information. In addition, the present work discusses problems encountered in the annotation and generation process.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.738}
}

@InProceedings{neves-coheur-nicolau:2020:LREC,
  author   = {Neves, Carolina and Coheur, Luísa and Nicolau, Hugo},
  title   = {HamNoSyS2SiGML: Translating HamNoSys Into SiGML},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6035--6039},
  abstract = {Sign Languages are visual languages and the main means of communication used by Deaf people. However, the majority of the information available online is presented through written form. Hence, it is not of easy access to the Deaf community. Avatars that can animate sign languages have gained an increase of interest in this area due to their flexibility in the process of generation and edition. Synthetic animation of conversational agents can be achieved through the use of notation systems. HamNoSys is one of these systems, which describes movements of the body through symbols. Its XML-compliant, SiGML, is a machine-readable input of HamNoSys able to animate avatars. Nevertheless, current tools have no freely available open source libraries that allow the conversion from HamNoSys to SiGML. Our goal is to develop a tool of open access, which can perform this conversion independently from other platforms. This system represents a crucial intermediate step in the

bigger pipeline of animating signing avatars. Two cases studies are described in order to illustrate different applications of our tool.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.739}
}

@InProceedings{kumar-EtAl:2020:LREC1,
  author   = {Kumar, Abhinav and Di Eugenio, Barbara and Aurisano, Jillian and Johnson, Andrew},
  title   = {Augmenting Small Data to Classify Contextualized Dialogue Acts for Exploratory Visualization},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {590--599},
  abstract = {Our goal is to develop an intelligent assistant to support users explore data via visualizations. We have collected a new corpus of conversations, CHICAGO-CRIME-VIS, geared towards supporting data visualization exploration, and we have annotated it for a variety of features, including contextualized dialogue acts. In this paper, we describe our strategies and their evaluation for dialogue act classification. We highlight how thinking aloud affects interpretation of dialogue acts in our setting and how to best capture that information. A key component of our strategy is data augmentation as applied to the training data, since our corpus is inherently small. We ran experiments with the Balanced Bagging Classifier (BAGC), Condiontal Random Field (CRF), and several Long Short Term Memory (LSTM) networks, and found that all of them improved compared to the baseline (e.g., without the data augmentation pipeline). CRF outperformed the other classification algorithms, with the LSTM networks showing modest improvement, even after obtaining a performance boost from domain-trained word embeddings. This result is of note because training a CRF is far less resource-intensive than training deep learning models, hence given a similar if not better performance, traditional methods may still be preferable in order to lower resource consumption.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.74}
}

@InProceedings{belissen-braffort-gouiffs:2020:LREC,
  author   = {Belissen, Valentin and Braffort, Annelies and Gouiffès, Michèle},
  title   = {Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {6040--6048},
  abstract  = {While the research in automatic Sign Language
Processing (SLP) is growing, it has been almost exclusively focused
on recognizing lexical signs, whether isolated or within continuous
SL production. However, Sign Languages include many other gestural
units like iconic structures, which need to be recognized in order
to go towards a true SL understanding. In this paper, we propose a
newer version of the publicly available SL corpus Dicta-Sign,
limited to its French Sign Language part. Involving 16 different
signers, this dialogue corpus was produced with very few constraints
on the style and content. It includes lexical and non-lexical
annotations over 11 hours of video recording, with 35000 manual
units. With the aim of stimulating research in SL understanding, we
also provide a baseline for the recognition of lexical signs and
non-lexical structures on this corpus. A very compact modeling of a
signer is built and a Convolutional-Recurrent Neural Network is
trained and tested on Dicta-Sign-LSF-v2, with state-of-the-art
results, including the ability to detect iconicity in SL
production.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.740}
}

@InProceedings{tornay-aran-magimaidoss:2020:LREC,
  author    = {Tornay, Sandrine  and  Aran, Oya  and  Magimai Doss,
Mathew},
  title     = {An HMM Approach with Inherent Model Selection for
Sign Language and Gesture Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6049--6056},
  abstract  = {HMMs have been the one of the first models to be
applied for sign recognition and have become the baseline models due
to their success in modeling sequential and multivariate data.
Despite the extensive use of HMMs for sign recognition, determining
the HMM structure has still remained as a challenge, especially when
the number of signs to be modeled is high. In this work, we present
a continuous HMM framework for modeling and recognizing isolated
signs, which inherently performs model selection to optimize the
number of states for each sign separately during recognition. Our
experiments on three different datasets, namely, German sign
language DGS dataset, Turkish sign language HospiSign dataset and
Chalearn14 dataset show that the proposed approach achieves better
sign language or gesture recognition systems in comparison to the
approach of selecting or presetting the number of HMM states based
on k-means, and yields systems that perform competitive to the case
where the number of states are determined based on the test set
performance.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.741}
}
```

```
@InProceedings{scicluna-strapparava:2020:LREC,
  author    = {Scicluna, Simone  and  Strapparava, Carlo},
  title     = {VROAV: Using Iconicity to Visually Represent Abstract
Verbs},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6057--6062},
  abstract  = {For a long time, philosophers, linguists and
scientists have been keen on finding an answer to the mind-bending
question “what does abstract language look like?", which has also
sprung from the phenomenon of mental imagery and how this emerges in
the mind. One way of approaching the matter of word representations
is by exploring the common semantic elements that link words to each
other. Visual languages like sign languages have been found to
reveal enlightening patterns across signs of similar meanings,
pointing towards the possibility of identifying clusters of iconic
meanings. With this insight, merged with an understanding of verb
predicates achieved from VerbNet, this study presents a novel verb
classification system based on visual shapes, using graphic
animation to visually represent 20 classes of abstract verbs.
Considerable agreement between participants who judged the graphic
animations based on representativeness suggests a positive way
forward for this proposal, which may be developed as a language
learning aid in educational contexts or as a multimodal language
comprehension tool for digital text.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.742}
}

@InProceedings{bull-braffort-gouiffs:2020:LREC,
  author    = {Bull, Hannah  and  Braffort, Annelies  and  Gouiffès,
Michèle},
  title     = {MEDIAPI-SKEL - A 2D-Skeleton Video Database of French
Sign Language With Aligned French Subtitles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6063--6068},
  abstract  = {This paper presents MEDIAPI-SKEL, a 2D-skeleton
database of French Sign Language videos aligned with French
subtitles. The corpus contains 27 hours of video of body, face and
hand keypoints, aligned to subtitles with a vocabulary size of 17k
tokens. In contrast to existing sign language corpora such as videos
produced under laboratory conditions or translations of TV programs
into sign language, this database is constructed using original sign
language content largely produced by deaf journalists at the media
company Média-Pi. Moreover, the videos are accurately synchronized
with French subtitles. We propose three challenges appropriate for
```

this corpus that are related to processing units of signs in context: automatic alignment of text and video, semantic segmentation of sign language, and production of video-text embeddings for cross-modal retrieval. These challenges deviate from the classic task of identifying a limited number of lexical signs in a video stream.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.743}
}

@InProceedings{kaczmarek-filhol:2020:LREC,
  author   = {Kaczmarek, Marion and Filhol, Michael},
  title   = {Alignment Data base for a Sign Language Concordancer},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6069--6072},
  abstract = {This article deals with elaborating a data base of alignments of parallel Franch-LSF segments. This data base is meant to be searched using a concordancer which we are also designing. We wish to equip Sign Language translators with tools similar to those used in text-to-text translation. To do so, we need language resources to feed them. Already existing Sign Language corpora can be found, but do not match our needs: working around a Sign Language concordancer, the corpus must be a parallel one and provide various examples of vocabulary and grammatical construction. We started with a parallel corpus of 40 short news and 120 SL videos , which we aligned manually by segments of various length. We described the methodology we used, how we define our segments and alignments. The last part concerns how we hope to allow the data base to keep growing in a near future.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.744}
}

@InProceedings{mukushev-EtAl:2020:LREC,
  author   = {Mukushev, Medet and Sabyrov, Arman and Imashev, Alfarabi and Koishybay, Kenessary and Kimmelman, Vadim and Sandygulova, Anara},
  title   = {Evaluation of Manual and Non-manual Components for Sign Language Recognition},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6073--6078},
  abstract = {The motivation behind this work lies in the need to differentiate between similar signs that differ in non-manual components present in any sign. To this end, we recorded full sentences signed by five native signers and extracted 5200 isolated

```
sign samples of twenty frequently used signs in Kazakh-Russian Sign
Language (K-RSL), which have similar manual components but differ in
non-manual components (i.e. facial expressions, eyebrow height,
mouth, and head orientation). We conducted a series of evaluations
in order to investigate whether non-manual components would improve
sign's recognition accuracy. Among standard machine learning
approaches, Logistic Regression produced the best results, 78.2\% of
accuracy for dataset with 20 signs and 77.9\% of accuracy for
dataset with 2 classes (statement vs question). Dataset can be
downloaded from the following website: https://
krslproject.github.io/krsl20/},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.745}
}

@InProceedings{kagirov-EtAl:2020:LREC,
  author    = {Kagirov, Ildar  and  Ivanko, Denis  and  Ryumin,
Dmitry  and  Axyonov, Alexander  and  Karpov, Alexey},
  title     = {TheRuSLan: Database of Russian Sign Language},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6079--6085},
  abstract  = {In this paper, a new Russian sign language multimedia
database TheRuSLan is presented. The database includes lexical units
(single words and phrases) from Russian sign language within one
subject area, namely, "food products at the supermarket", and was
collected using MS Kinect 2.0 device including both FullHD video and
the depth map modes, which provides new opportunities for the
lexicographical description of the Russian sign language vocabulary
and enhances research in the field of automatic gesture recognition.
Russian sign language has an official status in Russia, and over
120,000 deaf people in Russia and its neighboring countries use it
as their first language. Russian sign language has no writing
system, is poorly described and belongs to the low-resource
languages. The authors formulate the basic principles of annotation
of sign words, based on the collected data, and reveal the content
of the collected database. In the future, the database will be
expanded and comprise more lexical units. The database is explicitly
made for the task of creating an automatic system for Russian sign
language recognition.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.746}
}

@InProceedings{oshikawa-qian-wang:2020:LREC,
  author    = {Oshikawa, Ray  and  Qian, Jing  and  Wang, William
Yang},
  title     = {A Survey on Natural Language Processing for Fake News
Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6086--6093},
  abstract = {Fake news detection is a critical yet challenging problem in Natural Language Processing (NLP). The rapid rise of social networking platforms has not only yielded a vast increase in information accessibility but has also accelerated the spread of fake news. Thus, the effect of fake news has been growing, sometimes extending to the offline world and threatening public safety. Given the massive amount of Web content, automatic fake news detection is a practical NLP problem useful to all online content providers, in order to reduce the human time and effort to detect and prevent the spread of fake news. In this paper, we describe the challenges involved in fake news detection and also describe related tasks. We systematically review and compare the task formulations, datasets and NLP solutions that have been developed for this task, and also discuss the potentials and limitations of them. Based on our insights, we outline promising research directions, including more fine-grained, detailed, fair, and practical detection models. We also highlight the difference between fake news detection and other related tasks, and the importance of NLP solutions for fake news detection.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.747}
}

@InProceedings{gao-EtAl:2020:LREC,
  author   = {Gao, Jie and Han, Sooji and Song, Xingyi and Ciravegna, Fabio},
  title   = {RP-DNN: A Tweet Level Propagation Context Based Deep Neural Networks for Early Rumor Detection in Social Media},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6094--6105},
  abstract = {Early rumor detection (ERD) on social media platform is very challenging when limited, incomplete and noisy information is available. Most of the existing methods have largely worked on event-level detection that requires the collection of posts relevant to a specific event and relied only on user-generated content. They are not appropriate to detect rumor sources in the very early stages, before an event unfolds and becomes widespread. In this paper, we address the task of ERD at the message level. We present a novel hybrid neural network architecture, which combines a task-specific character-based bidirectional language model and stacked Long Short-Term Memory (LSTM) networks to represent textual contents and social-temporal contexts of input source tweets, for modelling propagation patterns of rumors in the early stages of their development. We apply multi-layered attention models to jointly learn attentive context embeddings over multiple context inputs. Our experiments employ a stringent leave-one-out cross-validation (LOO-

CV) evaluation setup on seven publicly available real-life rumor event data sets. Our models achieve state-of-the-art(SoA) performance for detecting unseen rumors on large augmented data which covers more than 12 events and 2,967 rumors. An ablation study is conducted to understand the relative contribution of each component of our proposed model.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.748}
}

@InProceedings{chen-huang-chen:2020:LREC2,
  author   = {Chen, Chung-Chi and Huang, Hen-Hsen and Chen, Hsin-Hsi},
  title   = {Issues and Perspectives from 10,000 Annotated Financial Social Media Data},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6106--6110},
  abstract = {In this paper, we investigate the annotation of financial social media data from several angles. We present Fin-SoMe, a dataset with 10,000 labeled financial tweets annotated by experts from both the front desk and the middle desk in a bank's treasury. These annotated results reveal that (1) writer-labeled market sentiment may be a misleading label; (2) writer's sentiment and market sentiment of an investor may be different; (3) most financial tweets provide unfounded analysis results; and (4) almost no investors write down the gain/loss results for their positions, which would otherwise greatly facilitate detailed evaluation of their performance. Based on these results, we address various open problems and suggest possible directions for future work on financial social media data. We also provide an experiment on the key snippet extraction task to compare the performance of using a general sentiment dictionary and using the domain-specific dictionary. The results echo our findings from the experts' annotations.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.749}
}

@InProceedings{paetzel-karkada-manuvinakurike:2020:LREC,
  author   = {Paetzel, Maike and Karkada, Deepthi and Manuvinakurike, Ramesh},
  title   = {RDG-Map: A Multimodal Corpus of Pedagogical Human-Agent Spoken Interactions.},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {600--609},
  abstract = {This paper presents a multimodal corpus of 209 spoken

game dialogues between a human and a remote-controlled artificial agent. The interactions involve people collaborating with the agent to identify countries on the world map as quickly as possible, which allows studying rapid and spontaneous dialogue with complex anaphoras, disfluent utterances and incorrect descriptions. The corpus consists of two parts: 8 hours of game interactions have been collected with a virtual unembodied agent online and 26.8 hours have been recorded with a physically embodied robot in a research lab. In addition to spoken audio recordings available for both parts, camera recordings and skeleton-, facial expression- and eye-gaze tracking data have been collected for the lab-based part of the corpus. In this paper, we introduce the pedagogical reference resolution game (RDG-Map) and the characteristics of the corpus collected. We also present an annotation scheme we developed in order to study the dialogue strategies utilized by the players. Based on a subset of 330 minutes of interactions annotated so far, we discuss initial insights into these strategies as well as the potential of the corpus for future research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.75}
}

@InProceedings{santos-funabashi-paraboni:2020:LREC,
  author   = {Santos, Wesley and Funabashi, Amanda and Paraboni, Ivandré},
  title   = {Searching Brazilian Twitter for Signs of Mental Health Issues},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6111--6117},
  abstract = {Depression and related mental health issues are often reflected in the language employed by the individuals who suffer from these conditions and, accordingly, research in Natural Language Processing (NLP) and related fields have developed an increasing number of studies devoted to their recognition in social media text. Some of these studies have also attempted to go beyond recognition by focusing on the early signs of these illnesses, and by analysing the users' publication history over time to potentially prevent further harm. The two kinds of study are of course overlapping, and often make use of supervised machine learning methods based on annotated corpora. However, as in many other fields, existing resources are largely devoted to English NLP, and there is little support for these studies in under resourced languages.To bridge this gap, in this paper we describe the initial steps towards building a novel resource of this kind - a corpus intended to support both the recognition of mental health issues and the temporal analysis of these illnesses - in the Brazilian Portuguese language, and initial results of a number of experiments in text classification addressing both tasks.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.750}
}

```
@InProceedings{tigunova-EtAl:2020:LREC,
  author    = {Tigunova, Anna  and  Mirza, Paramita  and  Yates,
Andrew  and  Weikum, Gerhard},
  title     = {RedDust: a Large Reusable Dataset of Reddit User
Traits},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6118--6126},
  abstract  = {Social media is a rich source of assertions about
personal traits, such as "I am a doctor" or "my hobby is playing
tennis". Precisely identifying explicit assertions is difficult,
though, because of the users' highly varied vocabulary and language
expressions. Identifying personal traits from implicit assertions
like I've been at work treating patients all day is even more
challenging. This paper presents RedDust, a large-scale annotated
resource for user profiling for over 300k Reddit users across five
attributes: profession, hobby, family status, age,and gender. We
construct RedDust using a diverse set of high-precision patterns and
demonstrate its use as a resource for developing learning models to
deal with implicit assertions. RedDust consists of users' personal
traits, which are (attribute, value) pairs, along with users' post
ids, which may be used to retrieve the posts from a publicly
available crawl or from the Reddit API. We discuss the construction
of the resource and show interesting statistics and insights into
the data. We also compare different classifiers, which can be
learned from RedDust. To the best of our knowledge, RedDust is the
first annotated language resource about Reddit users at large scale.
We envision further use cases of RedDust for providing background
knowledge about user traits, to enhance personalized search and
recommendation as well as conversational agents.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.751}
}

@InProceedings{bick:2020:LREC2,
  author    = {Bick, Eckhard},
  title     = {An Annotated Social Media Corpus for German},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6127--6135},
  abstract  = {This paper presents the German Twitter section of a
large (2 billion word) bilingual Social Media corpus for Hate Speech
research, discussing the compilation, pseudonymization and
grammatical annotation of the corpus, as well as special linguistic
features and peculiarities encountered in the data. Among other
things, compounding, accidental and intentional orthographic
```

variation, gendering and the use of emoticons/emojis are addressed in a genre-specific fashion. We present the different layers of linguistic annotation (morphosyntactic, dependencies and semantic types) and explain how a general parser (GerGram) can be made to work on Social Media data, pointing out necessary adaptations and extensions. In an evaluation run on a random cross-section of tweets, the modified parser achieved F-scores of 97\% for morphology (fine-grained POS) and 92\% for syntax (labeled attachment score). Predictably, performance was twice as good in tweets with standard orthography than in tweets with spelling/casing irregularities or lack of sentence separation, the effect being more marked for morphology than for syntax.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.752}
}

@InProceedings{weller-seppi:2020:LREC,
  author   = {Weller, Orion and Seppi, Kevin},
  title   = {The rJokes Dataset: a Large Scale Humor Collection},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6136--6141},
  abstract = {Humor is a complicated language phenomenon that depends upon many factors, including topic, date, and recipient. Because of this variation, it can be hard to determine what exactly makes a joke humorous, leading to difficulties in joke identification and related tasks. Furthermore, current humor datasets are lacking in both joke variety and size, with almost all current datasets having less than 100k jokes. In order to alleviate this issue we compile a collection of over 550,000 jokes posted over an 11 year period on the Reddit r/Jokes subreddit (an online forum), providing a large scale humor dataset that can easily be used for a myriad of tasks. This dataset also provides quantitative metrics for the level of humor in each joke, as determined by subreddit user feedback. We explore this dataset through the years, examining basic statistics, most mentioned entities, and sentiment proportions. We also introduce this dataset as a task for future work, where models learn to predict the level of humor in a joke. On that task we provide strong state-of-the-art baseline models and show room for future improvement. We hope that this dataset will not only help those researching computational humor, but also help social scientists who seek to understand popular culture through humor.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.753}
}

@InProceedings{proisl-EtAl:2020:LREC,
  author   = {Proisl, Thomas and Dykes, Natalie and Heinrich, Philipp and Kabashi, Besim and Blombach, Andreas and Evert, Stefan},
  title   = {EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatization and Semantic Tagging to a German Web and CMC Corpus},

  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6142--6148},
  abstract = {The EmpiriST corpus (Beißwenger et al., 2016) is a manually tokenized and part-of-speech tagged corpus of approximately 23,000 tokens of German Web and CMC (computer-mediated communication) data. We extend the corpus with manually created annotation layers for word form normalization, lemmatization and lexical semantics. All annotations have been independently performed by multiple human annotators. We report inter-annotator agreements and results of baseline systems and state-of-the-art off-the-shelf tools.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.754}
}

@InProceedings{nakamura-levy-wang:2020:LREC,
  author   = {Nakamura, Kai and Levy, Sharon and Wang, William Yang},
  title   = {Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6149--6157},
  abstract = {Fake news has altered society in negative ways in politics and culture. It has adversely affected both online social network systems as well as offline communities and conversations. Using automatic machine learning classification models is an efficient way to combat the widespread dissemination of fake news. However, a lack of effective, comprehensive datasets has been a problem for fake news research and detection model development. Prior fake news datasets do not provide multimodal text and image data, metadata, comment data, and fine-grained fake news categorization at the scale and breadth of our dataset. We present Fakeddit, a novel multimodal dataset consisting of over 1 million samples from multiple categories of fake news. After being processed through several stages of review, the samples are labeled according to 2-way, 3-way, and 6-way classification categories through distant supervision. We construct hybrid text+image models and perform extensive experiments for multiple variations of classification, demonstrating the importance of the novel aspect of multimodality and fine-grained classification unique to Fakeddit.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.755}
}

@InProceedings{sanders-vandenbosch:2020:LREC,
  author   = {Sanders, Eric and van den Bosch, Antal},

```
  title     = {Optimising Twitter-based Political Election
Prediction with Relevance andSentiment Filters},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6158--6165},
  abstract  = {We study the relation between the number of mentions
of political parties in the last weeks before the elections and the
election results.In this paper we focus on the Dutch elections of
the parliament in 2012 and for the provinces (and the senate) in
2011 and 2015. With raw counts, without adaptations, we achieve a
mean absolute error (MAE) of 2.71\% for 2011, 2.02\% for 2012 and
2.89\% for 2015. A set of over 17,000 tweets containing political
party names were annotated by at least three annotators per tweet on
ten features denoting communicative intent (including the presence
of sarcasm, the message's polarity, the presence of an explicit
voting endorsement or explicit voting advice, etc.). The annotations
were used to create oracle (gold-standard) filters. Tweets with or
without a certain majority annotation are held out from the tweet
counts, with the goal of attaining lower MAEs. With a grid search we
tested all combinations of filters and their responding MAE to find
the best filter ensemble. It appeared that the filters show markedly
different behaviour for the three elections and only a small MAE
improvement is possible when optimizing on all three elections.
Larger improvements for one election are possible, but result in
deterioration of the MAE for the other elections.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.756}
}

@InProceedings{iftene-EtAl:2020:LREC,
  author    = {Iftene, Adrian  and  Gifu, Daniela  and  Miron,
Andrei-Remus  and  Dudu, Mihai-Stefan},
  title     = {A Real-Time System for Credibility on Twitter},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6166--6173},
  abstract  = {Nowadays, social media credibility is a pressing
issue for each of us who are living in an altered online landscape.
The speed of news diffusion is striking. Given the popularity of
social networks, more and more users began posting pictures,
information, and news about personal life. At the same time, they
started to use all this information to get informed about what their
friends do or what is happening in the world, many of them arousing
much suspicion. The problem we are currently experiencing is that we
do not currently have an automatic method of figuring out in real-
time which news or which users are credible and which are not, what
is false or what is true on the Internet. The goal of this is to
```

analyze Twitter in real-time using neural networks in order to provide us key elements about both the credibility of tweets and users who posted them. Thus, we make a real-time heatmap using information gathered from users to create overall images of the areas from which this fake news comes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.757}
}

@InProceedings{ltekin:2020:LREC,
  author   = {Çöltekin, Çağrı},
  title   = {A Corpus of Turkish Offensive Language on Social Media},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6174--6184},
  abstract = {This paper introduces a corpus of Turkish offensive language. To our knowledge, this is the first corpus of offensive language for Turkish. The corpus consists of randomly sampled micro-blog posts from Twitter. The annotation guidelines are based on a careful review of the annotation practices of recent efforts for other languages. The corpus contains 36 232 tweets sampled randomly from the Twitter stream during a period of 18 months between Apr 2018 to Sept 2019. We found approximately 19 \% of the tweets in the data contain some type of offensive language, which is further subcategorized based on the target of the offense. We describe the annotation process, discuss some interesting aspects of the data, and present results of automatically classifying the corpus using state-of-the-art text classification methods. The classifiers achieve 77.3 \% F1 score on identifying offensive tweets, 77.9 \% F1 score on determining whether a given offensive document is targeted or not, and 53.0 \% F1 score on classifying the targeted offensive documents into three subcategories.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.758}
}

@InProceedings{seiffe-EtAl:2020:LREC,
  author   = {Seiffe, Laura and Marten, Oliver and Mikhailov, Michael and Schmeier, Sven and Möller, Sebastian and Roller, Roland},
  title   = {From Witch’s Shot to Music Making Bones - Resources for Medical Laymen to Technical Language and Vice Versa},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6185--6192},
  abstract = {Many people share information in social media or forums, like food they eat, sports activities they do or events

```
which have been visited. Information we share online unveil directly
or indirectly information about our lifestyle and health situation.
Particularly when text input is getting longer or multiple messages
can be linked to each other. Those information can be then used to
detect possible risk factors of diseases or adverse drug reactions
of medications. However, as most people are not medical experts,
language used might be more descriptive rather than the precise
medical expression as medics do. To detect and use those relevant
information, laymen language has to be translated and/or linked
against the corresponding medical concept. This work presents
baseline data sources in order to address this challenge for German
language. We introduce a new dataset which annotates medical laymen
and technical expressions in a patient forum, along with a set of
medical synonyms and definitions, and present first baseline results
on the data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.759}
}

@InProceedings{chen-huang-chen:2020:LREC1,
  author    = {Chen, Yi-Ting  and  Huang, Hen-Hsen  and  Chen, Hsin-
Hsi},
  title     = {MPDD: A Multi-Party Dialogue Dataset for Analysis of
Emotions and Interpersonal Relationships},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {610--614},
  abstract  = {A dialogue dataset is an indispensable resource for
building a dialogue system. Additional information like emotions and
interpersonal relationships labeled on conversations enables the
system to capture the emotion flow of the participants in the
dialogue. However, there is no publicly available Chinese dialogue
dataset with emotion and relation labels. In this paper, we collect
the conversions from TV series scripts, and annotate emotion and
interpersonal relationship labels on each utterance. This dataset
contains 25,548 utterances from 4,142 dialogues. We also set up some
experiments to observe the effects of the responded utterance on the
current utterance, and the correlation between emotion and relation
types in emotion and relation classification tasks.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.76}
}

@InProceedings{caselli-EtAl:2020:LREC,
  author    = {Caselli, Tommaso  and  Basile, Valerio  and
Mitrović, Jelena  and  Kartoziya, Inga  and  Granitzer, Michael},
  title     = {I Feel Offended, Don’t Be Abusive! Implicit/Explicit
Messages in Offensive and Abusive Language},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6193--6202},
  abstract  = {Abusive language detection is an unsolved and
challenging problem for the NLP community. Recent literature
suggests various approaches to distinguish between different
language phenomena (e.g., hate speech vs. cyberbullying vs.
offensive language) and factors (degree of explicitness and target)
that may help to classify different abusive language phenomena.
There are data sets that annotate the target of abusive messages
(i.e.OLID/OffensEval (Zampieri et al., 2019a)). However, there is a
lack of data sets that take into account the degree of explicitness.
In this paper, we propose annotation guidelines to distinguish
between explicit and implicit abuse in English and apply them to
OLID/OffensEval. The outcome is a newly created resource, AbuseEval
v1.0, which aims to address some of the existing issues in the
annotation of offensive and abusive language (e.g., explicitness of
the message, presence of a target, need of context, and interaction
across different phenomena).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.760}
}

@InProceedings{chowdhury-EtAl:2020:LREC,
  author    = {Chowdhury, Shammur Absar  and  Mubarak, Hamdy  and
Abdelali, Ahmed  and  Jung, Soon-gyo  and  Jansen, Bernard J  and
Salminen, Joni},
  title     = {A Multi-Platform Arabic News Comment Dataset for
Offensive Language Detection},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6203--6212},
  abstract  = {Access to social media often enables users to engage
in conversation with limited accountability. This allows a user to
share their opinions and ideology, especially regarding public
content, occasionally adopting offensive language. This may
encourage hate crimes or cause mental harm to targeted individuals
or groups. Hence, it is important to detect offensive comments in
social media platforms. Typically, most studies focus on offensive
commenting in one platform only, even though the problem of
offensive language is observed across multiple platforms. Therefore,
in this paper, we introduce and make publicly available a new
dialectal Arabic news comment dataset, collected from multiple
social media platforms, including Twitter, Facebook, and YouTube. We
follow two-step crowd-annotator selection criteria for low-
representative language annotation task in a crowdsourcing platform.
Furthermore, we analyze the distinctive lexical content along with
the use of emojis in offensive comments. We train and evaluate the
classifiers using the annotated multi-platform dataset along with
other publicly available data. Our results highlight the importance
of multiple platform dataset for (a) cross-platform, (b) cross-
```

domain, and (c) cross-dialect generalization of classifier performance.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.761}
}

@InProceedings{majdabadi-EtAl:2020:LREC,
  author   = {Majdabadi, Zahra and Sabeti, Behnam and Golazizian, Preni and Ashrafi Asli, Seyed Arad and Momenzadeh, Omid and fahmi, reza},
  title   = {Twitter Trend Extraction: A Graph-based Approach for Tweet and Hashtag Ranking, Utilizing No-Hashtag Tweets},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6213--6219},
  abstract = {Twitter has become a major platform for users to express their opinions on any topic and engage in debates. User debates and interactions usually lead to massive content regarding a specific topic which is called a Trend. Twitter trend extraction aims at finding these relevant groups of content that are generated in a short period. The most straightforward approach for this problem is using Hashtags, however, tweets without hashtags are not considered this way. In order to overcome this issue and extract trends using all tweets, we propose a graph-based approach where graph nodes represent tweets as well as words and hashtags. More specifically, we propose a modified version of RankClus algorithm to extract trends from the constructed tweets graph. The proposed approach is also capable of ranking tweets, words and hashtags in each trend with respect to their importance and relevance to the topic. The proposed algorithm is used to extract trends from several twitter datasets, where it produced consistent and coherent results.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.762}
}

@InProceedings{mazoyer-EtAl:2020:LREC,
  author   = {Mazoyer, Béatrice and Cagé, Julia and Hervé, Nicolas and Hudelot, Céline},
  title   = {A French Corpus for Event Detection on Twitter},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6220--6227},
  abstract = {We present Event2018, a corpus annotated for event detection tasks, consisting of 38 million tweets in French (retweets excluded) including more than 130,000 tweets manually annotated by three annotators as related or unrelated to a given event. The 243 events were selected both from press articles and from subjects

trending on Twitter during the annotation period (July to August 2018). In total, more than 95,000 tweets were annotated as related to one of the selected events. We also provide the titles and URLs of 15,500 news articles automatically detected as related to these events. In addition to this corpus, we detail the results of our event detection experiments on both this dataset and another publicly available dataset of tweets in English. We ran extensive tests with different types of text embeddings and a standard Topic Detection and Tracking algorithm, and detail our evaluation method. We show that tf-idf vectors allow the best performance for this task on both corpora. These results are intended to serve as a baseline for researchers wishing to test their own event detection systems on our corpus.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.763}
}

@InProceedings{chatterjere-EtAl:2020:LREC,
  author   = {Chatterjere, Arindam and Guptha, Vineeth and Chopra, Parul and Das, Amitava},
  title   = {Minority Positive Sampling for Switching Points - an Anecdote for the Code-Mixing Language Modeling},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6228--6236},
  abstract = {Code-Mixing (CM) or language mixing is a social norm in multilingual societies. CM is quite prevalent in social media conversations in multilingual regions like - India, Europe, Canada and Mexico. In this paper, we explore the problem of Language Modeling (LM) for code-mixed Hinglish text. In recent times, there have been several success stories with neural language modeling like Generative Pre-trained Transformer (GPT) (Radford et al., 2019), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) etc.. Hence, neural language models have become the new holy grail of modern NLP, although LM for CM is an unexplored area altogether. To better understand the problem of LM for CM, we initially experimented with several statistical language modeling techniques and consequently experimented with contemporary neural language models. Analysis shows switching-points are the main challenge for the LMCM performance drop, therefore in this paper we introduce the idea of minority positive sampling to selectively induce more sample to achieve better performance. On the contrary, all neural language models demand a huge corpus to train on for better performance. Finally, we are reporting a perplexity of 139 for Hinglish (Hindi-English language pair) LMCM using statistical bi-directional techniques.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.764}
}

@InProceedings{pamungkas-basile-patti:2020:LREC,
  author   = {Pamungkas, Endang Wahyu and Basile, Valerio and

Patti, Viviana},
  title   = {Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6237--6246},
  abstract = {Swearing plays an ubiquitous role in everyday conversations among humans, both in oral and textual communication, and occurs frequently in social media texts, typically featured by informal language and spontaneous writing. Such occurrences can be linked to an abusive context, when they contribute to the expression of hatred and to the abusive effect, causing harm and offense. However, swearing is multifaceted and is often used in casual contexts, also with positive social functions. In this study, we explore the phenomenon of swearing in Twitter conversations, taking the possibility of predicting the abusiveness of a swear word in a tweet context as the main investigation perspective. We developed the Twitter English corpus SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level. Our collection consists of 1,511 unique swear words from 1,320 tweets. We developed models to automatically predict abusive swearing, to provide an intrinsic evaluation of SWAD and confirm the robustness of the resource. We also present the results of a glass box ablation study in order to investigate which lexical, syntactic, and affective features are more informative towards the automatic prediction of the function of swearing.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.765}
}

@InProceedings{miao-last-litvak:2020:LREC,
  author   = {Miao, Lin and Last, Mark and Litvak, Marina},
  title   = {Detecting Troll Tweets in a Bilingual Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6247--6254},
  abstract = {During the past several years, a large amount of troll accounts has emerged with efforts to manipulate public opinion on social network sites. They are often involved in spreading misinformation, fake news, and propaganda with the intent of distracting and sowing discord. This paper aims to detect troll tweets in both English and Russian assuming that the tweets are generated by some "troll farm." We reduce this task to the authorship verification problem of determining whether a single tweet is authored by a "troll farm" account or not. We evaluate a supervised classification approach with monolingual, cross-lingual, and bilingual training scenarios, using several machine learning

algorithms, including deep learning. The best results are attained by the bilingual learning, showing the area under the ROC curve (AUC) of 0.875 and 0.828, for tweet classification in English and Russian test sets, respectively. It is noteworthy that these results are obtained using only raw text features, which do not require manual feature engineering efforts. In this paper, we introduce a resource of English and Russian troll tweets containing original tweets and translation from English to Russian, Russian to English. It is available for academic purposes.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.766}
}

@InProceedings{miletic-przewoznydesriaux-tanguy:2020:LREC,
  author   = {Miletic, Filip and Przewozny-Desriaux, Anne and Tanguy, Ludovic},
  title   = {Collecting Tweets to Investigate Regional Variation in Canadian English},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6255--6264},
  abstract = {We present a 78.8-million-tweet, 1.3-billion-word corpus aimed at studying regional variation in Canadian English with a specific focus on the dialect regions of Toronto, Montreal, and Vancouver. Our data collection and filtering pipeline reflects complex design criteria, which aim to allow for both data-intensive modeling methods and user-level variationist sociolinguistic analysis. It specifically consists in identifying Twitter users from the three cities, crawling their entire timelines, filtering the collected data in terms of user location and tweet language, and automatically excluding near-duplicate content. The resulting corpus mirrors national and regional specificities of Canadian English, it provides sufficient aggregate and user-level data, and it maintains a reasonably balanced distribution of content across regions and users. The utility of this dataset is illustrated by two example applications: the detection of regional lexical and topical variation, and the identification of contact-induced semantic shifts using vector space models. In accordance with Twitter's developer policy, the corpus will be publicly released in the form of tweet IDs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.767}
}

@InProceedings{abbes-EtAl:2020:LREC,
  author   = {Abbes, Ines and Zaghouani, Wajdi and El-Hardlo, Omaima and Ashour, Faten},
  title   = {DAICT: A Dialectal Arabic Irony Corpus Extracted from Twitter},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {6265--6271},
  abstract  = {Identifying irony in user-generated social media
content has a wide range of applications; however to date Arabic
content has received limited attention. To bridge this gap, this
study builds a new open domain Arabic corpus annotated for irony
detection. We query Twitter using irony-related hashtags to collect
ironic messages, which are then manually annotated by two linguists
according to our working definition of irony. Challenges which we
have encountered during the annotation process reflect the inherent
limitations of Twitter messages interpretation, as well as the
complexity of Arabic and its dialects. Once published, our corpus
will be a valuable free resource for developing open domain systems
for automatic irony recognition in Arabic language and its dialects
in social media text.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.768}
}

@InProceedings{vandergoot-EtAl:2020:LREC,
  author    = {van der Goot, Rob  and  Ramponi, Alan  and  Caselli,
Tommaso  and  Cafagna, Michele  and  De Mattei, Lorenzo},
  title     = {Norm It! Lexical Normalization for Italian and Its
Downstream Effects for Dependency Parsing},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6272--6278},
  abstract  = {Lexical normalization is the task of translating non-
standard social media data to a standard form. Previous work has
shown that this is beneficial for many downstream tasks in multiple
languages. However, for Italian, there is no benchmark available for
lexical normalization, despite the presence of many benchmarks for
other tasks involving social media data. In this paper, we discuss
the creation of a lexical normalization dataset for Italian. After
two rounds of annotation, a Cohen’s kappa score of 78.64 is
obtained. During this process, we also analyze the inter-annotator
agreement for this task, which is only rarely done on datasets for
lexical normalization,and when it is reported, the analysis usually
remains shallow. Furthermore, we utilize this dataset to train a
lexical normalization model and show that it can be used to improve
dependency parsing of social media data. All annotated data and the
code to reproduce the results are available at: http://
bitbucket.org/robvanderg/normit.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.769}
}

@InProceedings{siegert:2020:LREC,
  author    = {Siegert, Ingo},
  title     = {“Alexa in the wild” – Collecting Unconstrained
```

```
Conversations with a Modern Voice Assistant in a Public
Environment},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {615--619},
  abstract  = {Datasets featuring modern voice assistants such as
Alexa, Siri, Cortana and others allow an easy study of human-machine
interactions. But data collections offering an unconstrained,
unscripted public interaction are quite rare. Many studies so far
have focused on private usage, short pre-defined task or specific
domains. This contribution presents a dataset providing a large
amount of unconstrained public interactions with a voice assistant.
Up to now around 40 hours of device directed utterances were
collected during a science exhibition touring through Germany. The
data recording was part of an exhibit that engages visitors to
interact with a commercial voice assistant system (Amazon's ALEXA),
but did not restrict them to a specific topic. A specifically
developed quiz was starting point of the conversation, as the voice
assistant was presented to the visitors as a possible joker for the
quiz. But the visitors were not forced to solve the quiz with the
help of the voice assistant and thus many visitors had an open
conversation. The provided dataset – Voice Assistant Conversations
in the wild (VACW) – includes the transcripts of both visitors
requests and Alexa answers, identified topics and sessions as well
as acoustic characteristics automatically extractable from the
visitors' audio files.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.77}
}

@InProceedings{gugliotta-dinarelli:2020:LREC,
  author    = {Gugliotta, Elisa  and  Dinarelli, Marco},
  title     = {TArC: Incrementally and Semi-Automatically Collecting
a Tunisian Arabish Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6279--6286},
  abstract  = {This article describes the constitution process of
the first morpho-syntactically annotated Tunisian Arabish Corpus
(TArC). Arabish, also known as Arabizi, is a spontaneous coding of
Arabic dialects in Latin characters and "arithmographs" (numbers
used as letters). This code-system was developed by Arabic-speaking
users of social media in order to facilitate the writing in the
Computer-Mediated Communication (CMC) and text messaging informal
frameworks. Arabish differs for each Arabic dialect and each Arabish
code-system is under-resourced, in the same way as most of the
Arabic dialects. In the last few years, the attention of NLP studies
```

on Arabic dialects has considerably increased. Taking this into consideration, TArC will be a useful support for different types of analyses, computational and linguistic, as well as for NLP tools training. In this article we will describe preliminary work on the TArC semi-automatic construction process and some of the first analyses we developed on TArC. In addition, in order to provide a complete overview of the challenges faced during the building process, we will present the main Tunisian dialect characteristics and its encoding in Tunisian Arabish.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.770}
}

@InProceedings{rechkemmer-wilson-mihalcea:2020:LREC,
  author   = {Rechkemmer, Amy and Wilson, Steven and Mihalcea, Rada},
  title   = {Small Town or Metropolis? Analyzing the Relationship between Population Size and Language},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6287--6291},
  abstract = {The variance in language used by different cultures has been a topic of study for researchers in linguistics and psychology, but often times, language is compared across multiple countries in order to show a difference in culture. As a geographically large country that is diverse in population in terms of the background and experiences of its citizens, the U.S. also contains cultural differences within its own borders. Using a set of over 2 million posts from distinct Twitter users around the country dating back as far as 2014, we ask the following question: is there a difference in how Americans express themselves online depending on whether they reside in an urban or rural area? We categorize Twitter users as either urban or rural and identify ideas and language that are more commonly expressed in tweets written by one population over the other. We take this further by analyzing how the language from specific cities of the U.S. compares to the language of other cities and by training predictive models to predict whether a user is from an urban or rural area. We publicly release the tweet and user IDs that can be used to reconstruct the dataset for future studies in this direction.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.771}
}

@InProceedings{xu-prezrosas-mihalcea:2020:LREC,
  author   = {Xu, Zhentao and Pérez-Rosas, Verónica and Mihalcea, Rada},
  title   = {Inferring Social Media Users' Mental Health Status from Multimodal Information},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},

```
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6292--6299},
  abstract  = {Worldwide, an increasing number of people are
suffering from mental health disorders such as depression and
anxiety. In the United States alone, one in every four adults
suffers from a mental health condition, which makes mental health a
pressing concern. In this paper, we explore the use of multimodal
cues present in social media posts to predict users' mental health
status. Specifically, we focus on identifying social media activity
that either indicates a mental health condition or its onset. We
collect posts from Flickr and apply a multimodal approach that
consists of jointly analyzing language, visual, and metadata cues
and their relation to mental health. We conduct several
classification experiments aiming to discriminate between (1)
healthy users and users affected by a mental health illness; and (2)
healthy users and users prone to mental illness. Our experimental
results indicate that using multiple modalities can improve the
performance of this classification task as compared to the use of
one modality at a time, and can provide important cues into a user's
mental status.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.772}
}

@InProceedings{dekker-vandergoot:2020:LREC,
  author    = {Dekker, Kelly  and  van der Goot, Rob},
  title     = {Synthetic Data for English Lexical Normalization: How
Close Can We Get to Manually Annotated Data?},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6300--6309},
  abstract  = {Social media is a valuable data resource for various
natural language processing (NLP) tasks. However, standard NLP tools
were often designed with standard texts in mind, and their
performance decreases heavily when applied to social media data. One
solution to this problem is to adapt the input text to a more
standard form, a task also referred to as normalization. Automatic
approaches to normalization have shown that they can be used to
improve performance on a variety of NLP tasks. However, all of these
systems are supervised, thereby being heavily dependent on the
availability of training data for the correct language and domain.
In this work, we attempt to overcome this dependence by
automatically generating training data for lexical normalization.
Starting with raw tweets, we attempt two directions, to insert non-
standardness (noise) and to automatically normalize in an
unsupervised setting. Our best results are achieved by automatically
inserting noise. We evaluate our approaches by using an existing
lexical normalization system; our best scores are achieved by custom
error generation system, which makes use of some manually created
```

datasets. With this system, we score 94.29 accuracy on the test data, compared to 95.22 when it is trained on human-annotated data. Our best system which does not depend on any type of annotation is based on word embeddings and scores 92.04 accuracy. Finally, we perform an experiment in which we asked humans to predict whether a sentence was written by a human or generated by our best model. This experiment showed that in most cases it is hard for a human to detect automatically generated sentences.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.773}
}

@InProceedings{blombach-EtAl:2020:LREC,
  author   = {Blombach, Andreas and Dykes, Natalie and Heinrich, Philipp and Kabashi, Besim and Proisl, Thomas},
  title   = {A Corpus of German Reddit Exchanges (GeRedE)},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6310--6316},
  abstract = {GeRedE is a 270 million token German CMC corpus containing approximately 380,000 submissions and 6,800,000 comments posted on Reddit between 2010 and 2018. Reddit is a popular online platform combining social news aggregation, discussion and micro-blogging. Starting from a large, freely available data set, the paper describes our approach to filter out German data and further pre-processing steps, as well as which metadata and annotation layers have been included so far. We explore the Reddit sphere, what makes the German data linguistically peculiar, and how some of the communities within Reddit differ from one another. The CWB-indexed version of our final corpus is available via CQPweb, and all our processing scripts as well as all manual annotation and automatic language classification can be downloaded from GitHub.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.774}
}

@InProceedings{evrard-EtAl:2020:LREC,
  author   = {Evrard, Marc and Uro, Rémi and Hervé, Nicolas and Mazoyer, Béatrice},
  title   = {French Tweet Corpus for Automatic Stance Detection},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6317--6322},
  abstract = {The automatic stance detection task consists in determining the attitude expressed in a text toward a target (text, claim, or entity). This is a typical intermediate task for the fake news detection or analysis, which is a considerably widespread and a particularly difficult issue to overcome. This work aims at the

creation of a human-annotated corpus for the automatic stance detection of tweets written in French. It exploits a corpus of tweets collected during July and August 2018. To the best of our knowledge, this is the first freely available stance annotated tweet corpus in the French language. The four classes broadly adopted by the community were chosen for the annotation: support, deny, query, and comment with the addition of the ignore class. This paper presents the corpus along with the tools used to build it, its construction, an analysis of the inter-rater reliability, as well as the challenges and questions that were raised during the building process.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.775}
}

@InProceedings{abdikhojasteh-ansari-bohlouli:2020:LREC,
  author   = {Abdi Khojasteh, Hadi and Ansari, Ebrahim and Bohlouli, Mahdi},
  title   = {LSCP: Enhanced Large Scale Colloquial Persian Language Understanding},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6323--6327},
  abstract = {Language recognition has been significantly advanced in recent years by means of modern machine learning methods such as deep learning and benchmarks with rich annotations. However, research is still limited in low-resource formal languages. This consists of a significant gap in describing the colloquial language especially for low-resourced ones such as Persian. In order to target this gap for low resource languages, we propose a "Large Scale Colloquial Persian Dataset" (LSCP). LSCP is hierarchically organized in a semantic taxonomy that focuses on multi-task informal Persian language understanding as a comprehensive problem. This encompasses the recognition of multiple semantic aspects in the human-level sentences, which naturally captures from the real-world sentences. We believe that further investigations and processing, as well as the application of novel algorithms and methods, can strengthen enriching computerized understanding and processing of low resource languages. The proposed corpus consists of 120M sentences resulted from 27M tweets annotated with parsing tree, part-of-speech tags, sentiment polarity and translation in five different languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.776}
}

@InProceedings{oo-EtAl:2020:LREC,
  author   = {Oo, Yin May and Wattanavekin, Theeraphol and Li, Chenfang and De Silva, Pasindu and Sarin, Supheakmungkol and Pipatsrisawat, Knot and Jansche, Martin and Kjartansson, Oddur and Gutkin, Alexander},
  title   = {Burmese Speech Corpus, Finite-State Text

Normalization and Pronunciation Grammars with an Application to Text-to-Speech},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6328--6339},
  abstract = {This paper introduces an open-source crowd-sourced multi-speaker speech corpus along with the comprehensive set of finite-state transducer (FST) grammars for performing text normalization for the Burmese (Myanmar) language. We also introduce the open-source finite-state grammars for performing grapheme-to-phoneme (G2P) conversion for Burmese. These three components are necessary (but not sufficient) for building a high-quality text-to-speech (TTS) system for Burmese, a tonal Southeast Asian language from the Sino-Tibetan family which presents several linguistic challenges. We describe the corpus acquisition process and provide the details of our finite state-based approach to Burmese text normalization and G2P. Our experiments involve building a multi-speaker TTS system based on long short term memory (LSTM) recurrent neural network (RNN) models, which were previously shown to perform well for other languages in a low-resource setting. Our results indicate that the data and grammars that we are announcing are sufficient to build reasonably high-quality models comparable to other systems. We hope these resources will facilitate speech and language research on the Burmese language, which is considered by many to be low-resource due to the limited availability of free linguistic data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.777}
}

@InProceedings{booth-EtAl:2020:LREC2,
  author   = {Booth, Eric and Carns, Jake and Kennington, Casey and Rafla, Nader},
  title   = {Evaluating and Improving Child-Directed Automatic Speech Recognition},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6340--6345},
  abstract = {Speech recognition has seen dramatic improvements in the last decade, though those improvements have focused primarily on adult speech. In this paper, we assess child-directed speech recognition and leverage a transfer learning approach to improve child-directed speech recognition by training the recent DeepSpeech2 model on adult data, then apply additional tuning to varied amounts of child speech data. We evaluate our model using the CMU Kids dataset as well as our own recordings of child-directed prompts. The results from our experiment show that even a small amount of child

audio data improves significantly over a baseline of adult-only or child-only trained models. We report a final general Word-Error-Rate of 29\% over a baseline of 62\% that uses the adult-trained model. Our analyses show that our model adapts quickly using a small amount of data and that the general child model works better than school grade-specific models. We make available our trained model and our data collection tool.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.778}
}

@InProceedings{ihori-takashima-masumura:2020:LREC,
  author   = {Ihori, Mana and Takashima, Akihiko and Masumura, Ryo},
  title   = {Parallel Corpus for Japanese Spoken-to-Written Style Conversion},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6346--6353},
  abstract = {With the increase of automatic speech recognition (ASR) applications, spoken-to-written style conversion that transforms spoken-style text into written-style text is becoming an important technology to increase the readability of ASR transcriptions. To establish such conversion technology, a parallel corpus of spoken-style text and written-style text is beneficial because it can be utilized for building end-to-end neural sequence transformation models. Spoken-to-written style conversion involves multiple conversion problems including punctuation restoration, disfluency detection, and simplification. However, most existing corpora tend to be made for just one of these conversion problems. In addition, in Japanese, we have to consider not only general spoken-to-written style conversion problems but also Japanese-specific ones, such as language style unification (e.g., polite, frank, and direct styles) and omitted postpositional particle expressions restoration. Therefore, we created a new Japanese parallel corpus of spoken-style text and written-style text that can simultaneously handle general problems and Japanese-specific ones. To make this corpus, we prepared four types of spoken-style text and utilized a crowdsourcing service for manually converting them into written-style text. This paper describes the building setup of this corpus and reports the baseline results of spoken-to-written style conversion using the latest neural sequence transformation models.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.779}
}

@InProceedings{bothe-EtAl:2020:LREC,
  author   = {Bothe, Chandrakant and Weber, Cornelius and Magg, Sven and Wermter, Stefan},
  title   = {EDA: Enriching Emotional Dialogue Acts using an Ensemble of Neural Annotators},
  booktitle   = {Proceedings of The 12th Language Resources and

Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {620--627},
  abstract = {The recognition of emotion and dialogue acts enriches conversational analysis and help to build natural dialogue systems. Emotion interpretation makes us understand feelings and dialogue acts reflect the intentions and performative functions in the utterances. However, most of the textual and multi-modal conversational emotion corpora contain only emotion labels but not dialogue acts. To address this problem, we propose to use a pool of various recurrent neural models trained on a dialogue act corpus, with and without context. These neural models annotate the emotion corpora with dialogue act labels, and an ensemble annotator extracts the final dialogue act label. We annotated two accessible multi-modal emotion corpora: IEMOCAP and MELD. We analyzed the co-occurrence of emotion and dialogue act labels and discovered specific relations. For example, Accept/Agree dialogue acts often occur with the Joy emotion, Apology with Sadness, and Thanking with Joy. We make the Emotional Dialogue Acts (EDA) corpus publicly available to the research community for further study and analysis.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.78}
}

@InProceedings{gref-EtAl:2020:LREC,
  author   = {Gref, Michael and Walter, Oliver and Schmidt, Christoph and Behnke, Sven and Köhler, Joachim},
  title   = {Multi-Staged Cross-Lingual Acoustic Model Adaption for Robust Speech Recognition in Real-World Applications - A Case Study on German Oral History Interviews},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6354--6362},
  abstract = {While recent automatic speech recognition systems achieve remarkable performance when large amounts of adequate, high quality annotated speech data is used for training, the same systems often only achieve an unsatisfactory result for tasks in domains that greatly deviate from the conditions represented by the training data. For many real-world applications, there is a lack of sufficient data that can be directly used for training robust speech recognition systems. To address this issue, we propose and investigate an approach that performs a robust acoustic model adaption to a target domain in a cross-lingual, multi-staged manner. Our approach enables the exploitation of large-scale training data from other domains in both the same and other languages. We evaluate our approach using the challenging task of German oral history interviews, where we achieve a relative reduction of the word error

```
rate by more than 30\% compared to a model trained from scratch only
on the target domain, and 6-7\% relative compared to a model trained
robustly on 1000 hours of same-language out-of-domain training
data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.780}
}

@InProceedings{kratochvil-polak-bojar:2020:LREC,
  author    = {Kratochvil, Jonas  and  Polak, Peter  and  Bojar,
Ondrej},
  title     = {Large Corpus of Czech Parliament Plenary Hearings},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6363--6367},
  abstract  = {We present a large corpus of Czech parliament plenary
sessions. The corpus consists of approximately 1200 hours of speech
data and corresponding text transcriptions. The whole corpus has
been segmented to short audio segments making it suitable for both
training and evaluation of automatic speech recognition (ASR)
systems. The source language of the corpus is Czech, which makes it
a valuable resource for future research as only a few public
datasets are available in the Czech language. We complement the data
release with experiments of two baseline ASR systems trained on the
presented data: the more traditional approach implemented in the
Kaldi ASRtoolkit which combines hidden Markov models and deep neural
networks (NN) and a modern ASR architecture implemented in
Jaspertoolkit which uses deep NNs in an end-to-end fashion.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.781}
}

@InProceedings{szekely-edlund-gustafson:2020:LREC,
  author    = {Szekely, Eva  and  Edlund, Jens  and  gustafson,
joakim},
  title     = {Augmented Prompt Selection for Evaluation of
Spontaneous Speech Synthesis},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6368--6374},
  abstract  = {By definition, spontaneous speech is unscripted and
created on the fly by the speaker. It is dramatically different from
read speech, where the words are authored as text before they are
spoken. Spontaneous speech is emergent and transient, whereas text
read out loud is pre-planned. For this reason, it is unsuitable to
evaluate the usability and appropriateness of spontaneous speech
synthesis by having it read out written texts sampled from for
example newspapers or books. Instead, we need to use transcriptions
```

of speech as the target - something that is much less readily available. In this paper, we introduce Starmap, a tool allowing developers to select a varied, representative set of utterances from a spoken genre, to be used for evaluation of TTS for a given domain. The selection can be done from any speech recording, without the need for transcription. The tool uses interactive visualisation of prosodic features with t-SNE, along with a tree-based algorithm to guide the user through thousands of utterances and ensure coverage of a variety of prompts. A listening test has shown that with a selection of genre-specific utterances, it is possible to show significant differences across genres between two synthetic voices built from spontaneous speech.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.782}
}

@InProceedings{schulder-EtAl:2020:LREC,
  author   = {Schulder, Marc and O'Mahony, Johannah and Bakanouski, Yury and Klakow, Dietrich},
  title   = {ATC-ANNO: Semantic Annotation for Air Traffic Control with Assistive Auto-Annotation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6375--6380},
  abstract = {In air traffic control, assistant systems support air traffic controllers in their work. To improve the reactivity and accuracy of the assistant, automatic speech recognition can monitor the commands uttered by the controller. However, to provide sufficient training data for the speech recognition system, many hours of air traffic communications have to be transcribed and semantically annotated. For this purpose we developed the annotation tool ATC-ANNO. It provides a number of features to support the annotator in their task, such as auto-complete suggestions for semantic tags, access to preliminary speech recognition predictions, syntax highlighting and consistency indicators. Its core assistive feature, however, is its ability to automatically generate semantic annotations. Although it is based on a simple hand-written finite state grammar, it is also able to annotate sentences that deviate from this grammar. We evaluate the impact of different features on annotator efficiency and find that automatic annotation allows annotators to cover four times as many utterances in the same time.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.783}
}

@InProceedings{hernandezmena-EtAl:2020:LREC,
  author   = {Hernandez Mena, Carlos Daniel and Gatt, Albert and DeMarco, Andrea and Borg, Claudia and van der Plas, Lonneke and Muscat, Amanda and Padovani, Ian},
  title   = {MASRI-HEADSET: A Maltese Corpus for Speech Recognition},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6381--6388},
  abstract  = {Maltese, the national language of Malta, is spoken by
approximately 500,000 people. Speech processing for Maltese is still
in its early stages of development. In this paper, we present the
first spoken Maltese corpus designed purposely for Automatic Speech
Recognition (ASR). The MASRI-HEADSET corpus was developed by the
MASRI project at the University of Malta. It consists of 8 hours of
speech paired with text, recorded by using short text snippets in a
laboratory environment. The speakers were recruited from different
geographical locations all over the Maltese islands, and were
roughly evenly distributed by gender. This paper also presents some
initial results achieved in baseline experiments for Maltese ASR
using Sphinx and Kaldi. The MASRI HEADSET Corpus is publicly
available for research/academic purposes.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.784}
}

@InProceedings{kalashnikova-EtAl:2020:LREC,
  author    = {Kalashnikova, Natalia  and  Grobol, Loïc  and
Eshkol-Taravella, Iris  and  Delafontaine, François},
  title     = {Automatic Period Segmentation of Oral French},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6389--6394},
  abstract  = {Natural Language Processing in oral speech
segmentation is still looking for a minimal unit to analyze. In this
work, we present a comparison of two automatic segmentation methods
of macro-syntactic periods which allows to take into account
syntactic and prosodic components of speech. We compare the
performances of an existing tool Analor (Avanzi, Lacheret-Dujour,
Victorri, 2008) developed for automatic segmentation of prosodic
periods and of CRF models relying on syntactic and / or prosodic
features. We find that Analor tends to divide speech into smaller
segments and that CRF models detect larger segments rather than
macro-syntactic periods. However, in general CRF models perform
better results than Analor in terms of F-measure.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.785}
}

@InProceedings{desot-portet-vacher:2020:LREC,
  author    = {Desot, Thierry  and  Portet, François  and  Vacher,
Michel},
  title     = {Corpus Generation for Voice Command in Smart Home and
the Effect of Speech Synthesis on End-to-End SLU},
```

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6395--6404},
  abstract  = {Massive amounts of annotated data greatly contributed
to the advance of the machine learning field. However such large
data sets are often unavailable for novel tasks performed in
realistic environments such as smart homes. In this domain,
semantically annotated large voice command corpora for Spoken
Language Understanding (SLU) are scarce, especially for non-English
languages. We present the automatic generation process of a
synthetic semantically-annotated corpus of French commands for
smart-home to train pipeline and End-to-End (E2E) SLU models. SLU is
typically performed through Automatic Speech Recognition (ASR) and
Natural Language Understanding (NLU) in a pipeline. Since errors at
the ASR stage reduce the NLU performance, an alternative approach is
End-to-End (E2E) SLU to jointly perform ASR and NLU. To that end,
the artificial corpus was fed to a text-to-speech (TTS) system to
generate synthetic speech data. All models were evaluated on voice
commands acquired in a real smart home. We show that artificial data
can be combined with real data within the same training set or used
as a stand-alone training corpus. The synthetic speech quality was
assessedby comparing it to real data using dynamic time warping
(DTW).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.786}
}

@InProceedings{benabdallah-kchaou-bougares:2020:LREC,
  author    = {Ben Abdallah, Najla  and  Kchaou, Saméh  and
Bougares, Fethi},
  title     = {Text and Speech-based Tunisian Arabic Sub-Dialects
Identification},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6405--6411},
  abstract  = {Dialect IDentification (DID) is a challenging task,
and it becomes more complicated when it is about the identification
of dialects that belong to the same country. Indeed, dialects of the
same country are closely related and exhibit a significant
overlapping at the phonetic and lexical levels. In this paper, we
present our first results on a dialect classification task covering
four sub-dialects spoken in Tunisia. We use the term 'sub-dialect'
to refer to the dialects belonging to the same country. We conducted
our experiments aiming to discriminate between Tunisian sub-dialects
belonging to four different cities: namely Tunis, Sfax, Sousse and
Tataouine. A spoken corpus of 1673 utterances is collected,
transcribed and freely distributed. We used this corpus to build
```

```
several speech- and text-based DID systems. Our results confirm
that, at this level of granularity, dialects are much better
distinguishable using the speech modality. Indeed, we were able to
reach an F-1 score of 93.75\% using our best speech-based
identification system while the F-1 score is limited to 54.16\%
using text-based DID on the same test set.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.787}
}

@InProceedings{rognoni-EtAl:2020:LREC,
  author    = {Rognoni, Luca  and  Bishop, Judith  and  Corris,
Miriam  and  Fernando, Jessica  and  Smith, Rosanna},
  title     = {Urdu Pitch Accents and Intonation Patterns in
Spontaneous Conversational Speech},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6412--6416},
  abstract  = {An intonational inventory of Urdu for spontaneous
conversational speech is determined based on the analysis of a hand-
labelled data set of telephone conversations. An inventory of Urdu
pitch accents and the basic Urdu intonation patterns observed in the
data are summarised and presented using a simplified version of the
Rhythm and Pitch (RaP) labelling system. The relation between pitch
accents and parts of speech (PoS) is also explored. The data confirm
the important role played by low pitch accents in Urdu spontaneous
speech, in line with previous studies on Urdu/Hindi scripted speech.
Typical pitch contours such as falling tone in statements and WH-
questions, and rising tone for yes/no questions are also exhibited.
Pitch accent distribution is quite free in Urdu, but the data
indicate a stronger association of pitch accent with some PoS
categories of content word (e.g. Nouns) when compared with function
words and semantically lighter PoS categories (such as Light Verbs).
Contrastive focus is realised by an L*+H accent with a relatively
large pitch excursion for the +H tone, and longer duration of the
stressed syllable. The data suggest that post-focus compression
(PFC) is used in Urdu as a focus-marking strategy.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.788}
}

@InProceedings{srivastava-EtAl:2020:LREC,
  author    = {Srivastava, Nimisha  and  Mukhopadhyay, Rudrabha  and
K R, Prajwal  and  Jawahar, C V},
  title     = {IndicSpeech: Text-to-Speech Corpus for Indian
Languages},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {6417--6422},
  abstract  = {India is a country where several tens of languages
are spoken by over a billion strong population. Text-to-speech
systems for such languages will thus be extremely beneficial for
wide-spread content creation and accessibility. Despite this, the
current TTS systems for even the most popular Indian languages fall
short of the contemporary state-of-the-art systems for English,
Chinese, etc. We believe that one of the major reasons for this is
the lack of large, publicly available text-to-speech corpora in
these languages that are suitable for training neural text-to-speech
systems. To mitigate this, we release a $24$ hour text-to-speech
corpus for $3$ major Indian languages namely Hindi, Malayalam and
Bengali. In this work, we also train a state-of-the-art TTS system
for each of these languages and report their performances. The
collected corpus, code, and trained models are made publicly
available.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.789}
}

@InProceedings{amoyal-priegovalverde-rauzy:2020:LREC,
  author    = {Amoyal, Mary  and  Priego-Valverde, Béatrice  and
Rauzy, Stephane},
  title     = {PACO: a Corpus to Analyze the Impact of Common Ground
in Spontaneous Face-to-Face Interaction},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {628--633},
  abstract  = {PAC0 is a French audio-video conversational corpus
made of 15 face-to-face dyadic interactions, lasting around 20 min
each. This compared corpus has been created in order to explore the
impact of the lack of personal common ground (Clark, 1996) on
participants collaboration during conversation and specifically on
their smile during topic transitions. We have constituted this
conversational corpus " PACO" by replicating the experimental
protocol of "Cheese!" (Priego-valverde \& al.,2018). The only
difference that distinguishes these two corpora is the degree of CG
of the interlocutors: in Cheese! interlocutors are friends, while in
PACO they do not know each other. This experimental protocol allows
to analyze how the participants are getting acquainted. This study
brings two main contributions. First, the PACO conversational corpus
enables to compare the impact of the interlocutors' common ground.
Second, the semi-automatic smile annotation protocol allows to
obtain reliable and reproducible smile annotations while reducing
the annotation time by a factor 10. Keywords : Common ground,
spontaneous interaction, smile, automatic detection.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.79}
}

@InProceedings{gorisch-gref-schmidt:2020:LREC,
  author    = {Gorisch, Jan  and  Gref, Michael  and  Schmidt,
```

Thomas},
  title   = {Using Automatic Speech Recognition in Spoken Corpus Curation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6423--6428},
  abstract = {The newest generation of speech technology caused a huge increase of audio-visual data nowadays being enhanced with orthographic transcripts such as in automatic subtitling in online platforms. Research data centers and archives contain a range of new and historical data, which are currently only partially transcribed and therefore only partially accessible for systematic querying. Automatic Speech Recognition (ASR) is one option of making that data accessible. This paper tests the usability of a state-of-the-art ASR-System on a historical (from the 1960s), but regionally balanced corpus of spoken German, and a relatively new corpus (from 2012) recorded in a narrow area. We observed a regional bias of the ASR-System with higher recognition scores for the north of Germany vs. lower scores for the south. A detailed analysis of the narrow region data revealed -- despite relatively high ASR-confidence -- some specific word errors due to a lack of regional adaptation. These findings need to be considered in decisions on further data processing and the curation of corpora, e.g. correcting transcripts or transcribing from scratch. Such geography-dependent analyses can also have the potential for ASR-development to make targeted data selection for training/adaptation and to increase the sensitivity towards varieties of pluricentric languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.790}
}

@InProceedings{deng-EtAl:2020:LREC,
  author   = {Deng, Huaijin and Lin, Youchao and Utsuro, Takehito and Kobayashi, Akio and Nishizaki, Hiromitsu and Hoshino, Junichi},
  title   = {Integrating Disfluency-based and Prosodic Features with Acoustics in Automatic Fluency Evaluation of Spontaneous Speech},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6429--6437},
  abstract = {This paper describes an automatic fluency evaluation of spontaneous speech. In the task of automatic fluency evaluation, we integrate diverse features of acoustics, prosody, and disfluency-based ones. Then, we attempt to reveal the contribution of each of those diverse features to the task of automatic fluency evaluation. Although a variety of different disfluencies are observed regularly

```
in spontaneous speech, we focus on two types of phenomena, i.e., filled pauses and word fragments. The experimental results demonstrate that the disfluency-based features derived from word fragments and filled pauses are effective relative to evaluating fluent/disfluent speech, especially when combined with prosodic features, e.g., such as speech rate and pauses/silence. Next, we employed an LSTM based framework in order to integrate the disfluency-based and prosodic features with time sequential acoustic features. The experimental evaluation results of those integrated diverse features indicate that time sequential acoustic features contribute to improving the model with disfluency-based and prosodic features when detecting fluent speech, but not when detecting disfluent speech. Furthermore, when detecting disfluent speech, the model without time sequential acoustic features performs best even without word fragments features, but only with filled pauses and prosodic features.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.791}
}

@InProceedings{yamashita-EtAl:2020:LREC,
  author    = {Yamashita, Yuki  and  Koriyama, Tomoki  and  Saito, Yuki  and  Takamichi, Shinnosuke  and  Ijima, Yusuke  and  Masumura, Ryo  and  Saruwatari, Hiroshi},
  title     = {DNN-based Speech Synthesis Using Abundant Tags of Spontaneous Speech Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6438--6443},
  abstract  = {In this paper, we investigate the effectiveness of using rich annotations in deep neural network (DNN)-based statistical speech synthesis. DNN-based frameworks typically use linguistic information as input features called context instead of directly using text. In such frameworks, we can synthesize not only reading-style speech but also speech with paralinguistic and nonlinguistic features by adding such information to the context. However, it is not clear what kind of information is crucial for reproducing paralinguistic and nonlinguistic features. Therefore, we investigate the effectiveness of rich tags in DNN-based speech synthesis according to the Corpus of Spontaneous Japanese (CSJ), which has a large amount of annotations on paralinguistic features such as prosody, disfluency, and morphological features. Experimental evaluation results shows that the reproducibility of paralinguistic features of synthetic speech was enhanced by adding such information as context.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.792}
}

@InProceedings{abulimiti-schultz:2020:LREC,
  author    = {Abulimiti, Ayimunishagu  and  Schultz, Tanja},
  title     = {Automatic Speech Recognition for Uyghur through
```

```
Multilingual Acoustic Modeling},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6444--6449},
  abstract  = {Low-resource languages suffer from lower performance
of Automatic Speech Recognition (ASR) system due to the lack of
data. As a common approach, multilingual training has been applied
to achieve more context coverage and has shown better performance
over the monolingual training (Heigold et al., 2013). However, the
difference between the donor language and the target language may
distort the acoustic model trained with multilingual data,
especially when much larger amount of data from donor languages is
used for training the models of low-resource language. This paper
presents our effort towards improving the performance of ASR system
for the under-resourced Uyghur language with multilingual acoustic
training. For the developing of multilingual speech recognition
system for Uyghur, we used Turkish as donor language, which we
selected from GlobalPhone corpus as the most similar language to
Uyghur. By generating subsets of Uyghur training data, we explored
the performance of multilingual speech recognition systems trained
with different sizes of Uyghur and Turkish data. The best speech
recognition system for Uyghur is achieved by multilingual training
using all Uyghur data (10hours) and 17 hours of Turkish data and the
WER is 19.17\%, which corresponds to 4.95\% relative improvement
over monolingual training.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.793}
}

@InProceedings{delgado-EtAl:2020:LREC,
  author    = {Delgado, Dana  and  Walker, Kevin  and  Strassel,
Stephanie  and  Jones, Karen  and  Caruso, Christopher  and  Graff,
David},
  title     = {The SAFE-T Corpus: A New Resource for Simulated
Public Safety Communications},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6450--6457},
  abstract  = {We introduce a new resource, the SAFE-T (Speech
Analysis for Emergency Response Technology) Corpus, designed to
simulate first-responder communications by inducing high vocal
effort and urgent speech with situational background noise in a
game-based collection protocol. Linguistic Data Consortium developed
the SAFE-T Corpus to support the NIST (National Institute of
Standards and Technology) OpenSAT (Speech Analytic Technologies)
evaluation series, whose goal is to advance speech analytic
technologies including automatic speech recognition, speech activity
```

```
detection and keyword search in multiple domains including simulated
public safety communications data. The corpus comprises over 300
hours of audio from 115 unique speakers engaged in a collaborative
problem-solving activity representative of public safety
communications in terms of speech content, noise types and noise
levels. Portions of the corpus have been used in the OpenSAT 2019
evaluation and the full corpus will be published in the LDC catalog.
We describe the design and implementation of the SAFE-T Corpus
collection, discuss the approach of capturing spontaneous speech
from study participants through game-based speech collection, and
report on the collection results including several challenges
associated with the collection.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.794}
}

@InProceedings{gogoi-EtAl:2020:LREC,
  author    = {Gogoi, Parismita  and  Dey, Abhishek  and
Lalhminghlui, Wendy  and  Sarmah, Priyankoo  and  Prasanna, S R
Mahadeva},
  title     = {Lexical Tone Recognition in Mizo using Acoustic-
Prosodic Features},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6458--6461},
  abstract  = {Mizo is an under-studied Tibeto-Burman tonal language
of the North-East India. Preliminary research findings have
confirmed that four distinct tones of Mizo (High, Low, Rising and
Falling) appear in the language. In this work, an attempt is made to
automatically recognize four phonological tones in Mizo
distinctively using acoustic-prosodic parameters as features. Six
features computed from Fundamental Frequency (F0) contours are
considered and two classifier models based on Support Vector Machine
(SVM) \& Deep Neural Network (DNN) are implemented for automatic
tonerecognition task respectively. The Mizo database consists of
31950 iterations of the four Mizo tones, collected from 19 speakers
using trisyllabic phrases. A four-way classification of tones is
attempted with a balanced (equal number of iterations per tone
category) dataset for each tone of Mizo. it is observed that the DNN
based classifier shows comparable performance in correctly
recognizing four phonological Mizo tones as of the SVM based
classifier.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.795}
}

@InProceedings{meyer-EtAl:2020:LREC,
  author    = {Meyer, Josh  and  Rauchenstein, Lindy  and
Eisenberg, Joshua D.  and  Howell, Nicholas},
  title     = {Artie Bias Corpus: An Open Dataset for Detecting
Demographic Bias in Speech Applications},
  booktitle      = {Proceedings of The 12th Language Resources and
```

```
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6462--6468},
  abstract  = {We describe the creation of the Artie Bias Corpus, an
English dataset of expert-validated <audio, transcript> pairs with
demographic tags for {age, gender, accent}. We also release open
software which may be used with the Artie Bias Corpus to detect
demographic bias in Automatic Speech Recognition systems, and can be
extended to other speech technologies. The Artie Bias Corpus is a
curated subset of the Mozilla Common Voice corpus, which we release
under a Creative Commons CC0 license — the most open and permissive
license for data. This article contains information on the criteria
used to select and annotate the Artie Bias Corpus in addition to
experiments in which we detect and attempt to mitigate bias in end-
to-end speech recognition models. We we observe a significant accent
bias in our baseline DeepSpeech model, with more accurate
transcriptions of US English compared to Indian English. We do not,
however, find evidence for a significant gender bias. We then show
significant improvements on individual demographic groups from fine-
tuning.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.796}
}

@InProceedings{georgila-EtAl:2020:LREC2,
  author    = {Georgila, Kallirroi  and  Leuski, Anton  and  Yanov,
Volodymyr  and  Traum, David},
  title     = {Evaluation of Off-the-shelf Speech Recognizers Across
Diverse Dialogue Domains},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6469--6476},
  abstract  = {We evaluate several publicly available off-the-shelf
(commercial and research) automatic speech recognition (ASR) systems
across diverse dialogue domains (in US-English). Our evaluation is
aimed at non-experts with limited experience in speech recognition.
Our goal is not only to compare a variety of ASR systems on several
diverse data sets but also to measure how much ASR technology has
advanced since our previous large-scale evaluations on the same data
sets. Our results show that the performance of each speech
recognizer can vary significantly depending on the domain.
Furthermore, despite major recent progress in ASR technology,
current state-of-the-art speech recognizers perform poorly in
domains that require special vocabulary and language models, and
under noisy conditions. We expect that our evaluation will prove
useful to ASR consumers and dialogue system designers.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.797}
}
```

```
@InProceedings{ulasik-EtAl:2020:LREC,
  author    = {Ulasik, Malgorzata Anna  and  Hürlimann, Manuela  and
Germann, Fabian  and  Gedik, Esin  and  Benites, Fernando  and
Cieliebak, Mark},
  title     = {CEASR: A Corpus for Evaluating Automatic Speech
Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6477--6485},
  abstract  = {In this paper, we present CEASR, a Corpus for
Evaluating the quality of Automatic Speech Recognition (ASR). It is
a data set based on public speech corpora, containing metadata along
with transcripts generated by several modern state-of-the-art ASR
systems. CEASR provides this data in a unified structure, consistent
across all corpora and systems, with normalised transcript texts and
metadata. We use CEASR to evaluate the quality of ASR systems by
calculating an average Word Error Rate (WER) per corpus, per system
and per corpus-system pair. Our experiments show a substantial
difference in accuracy between commercial versus open-source ASR
tools as well as differences up to a factor ten for single systems
on different corpora. Using CEASR allowed us to very efficiently and
easily obtain these results. Our corpus enables researchers to
perform ASR-related evaluations and various in-depth analyses with
noticeably reduced effort, i.e. without the need to collect, process
and transcribe the speech data themselves.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.798}
}

@InProceedings{zanonboito-EtAl:2020:LREC,
  author    = {Zanon Boito, Marcely  and  Havard, William  and
Garnerin, Mahault  and  Le Ferrand, Éric  and  Besacier, Laurent},
  title     = {MaSS: A Large and Clean Multilingual Corpus of
Sentence-aligned Spoken Utterances Extracted from the Bible},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6486--6493},
  abstract  = {The CMU Wilderness Multilingual Speech Dataset
(Black, 2019) is a newly published multilingual speech dataset based
on recorded readings of the New Testament. It provides data to build
Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) models
for potentially 700 languages. However, the fact that the source
content (the Bible) is the same for all the languages is not
exploited to date.Therefore, this article proposes to add
multilingual links between speech segments in different languages,
and shares a large and clean dataset of 8,130 parallel spoken
```

utterances across 8 languages (56 language pairs). We name this corpus MaSS (Multilingual corpus of Sentence-aligned Spoken utterances). The covered languages (Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish) allow researches on speech-to-speech alignment as well as on translation for typologically different language pairs. The quality of the final corpus is attested by human evaluation performed on a corpus subset (100 utterances, 8 language pairs). Lastly, we showcase the usefulness of the final product on a bilingual speech retrieval task.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.799}
}

@InProceedings{dakle-desai-moldovan:2020:LREC,
  author   = {Dakle, Parag Pravin and Desai, Takshak and Moldovan, Dan},
  title   = {A Study on Entity Resolution for Email Conversations},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {65--73},
  abstract = {This paper investigates the problem of entity resolution for email conversations and presents a seed annotated corpus of email threads labeled with entity coreference chains. Characteristics of email threads concerning reference resolution are first discussed, and then the creation of the corpus and annotation steps are explained. Finally, performance of the current state-of-the-art deep learning models on the seed corpus is evaluated and qualitative error analysis on the predictions obtained is presented.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.8}
}

@InProceedings{navarretta-paggio:2020:LREC,
  author   = {Navarretta, Costanza and Paggio, Patrizia},
  title   = {Dialogue Act Annotation in a Multimodal Corpus of First Encounter Dialogues},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {634--643},
  abstract = {This paper deals with the annotation of dialogue acts in a multimodal corpus of first encounter dialogues, i.e. face-to-face dialogues in which two people who meet for the first time talk with no particular purpose other than just talking. More specifically, we describe the method used to annotate dialogue acts in the corpus, including the evaluation of the annotations. Then, we

present descriptive statistics of the annotation, particularly focusing on which dialogue acts often follow each other across speakers and which dialogue acts overlap with gestural behaviour. Finally, we discuss how feedback is expressed in the corpus by means of feedback dialogue acts with or without co-occurring gestural behaviour, i.e. multimodal vs. unimodal feedback.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.80}
}

@InProceedings{he-EtAl:2020:LREC,
  author   = {He, Fei and Chu, Shan-Hui Cathy and Kjartansson, Oddur and Rivera, Clara and Katanova, Anna and Gutkin, Alexander and Demirsahin, Isin and Johny, Cibu and Jansche, Martin and Sarin, Supheakmungkol and Pipatsrisawat, Knot},
  title   = {Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6494--6503},
  abstract = {We present free high quality multi-speaker speech corpora for Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu, which are six of the twenty two official languages of India spoken by 374 million native speakers. The datasets are primarily intended for use in text-to-speech (TTS) applications, such as constructing multilingual voices or being used for speaker or language adaptation. Most of the corpora (apart from Marathi, which is a female-only database) consist of at least 2,000 recorded lines from female and male native speakers of the language. We present the methodological details behind corpora acquisition, which can be scaled to acquiring data for other languages of interest. We describe the experiments in building a multilingual text-to-speech model that is constructed by combining our corpora. Our results indicate that using these corpora results in good quality voices, with Mean Opinion Scores (MOS) > 3.6, for all the languages tested. We believe that these resources, released with an open-source license, and the described methodology will help in the progress of speech applications for the languages described and aid corpora development for other, smaller, languages of India and beyond.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.800}
}

@InProceedings{guevararukoz-EtAl:2020:LREC,
  author   = {Guevara-Rukoz, Adriana and Demirsahin, Isin and He, Fei and Chu, Shan-Hui Cathy and Sarin, Supheakmungkol and Pipatsrisawat, Knot and Gutkin, Alexander and Butryna, Alena and Kjartansson, Oddur},
  title   = {Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech},
  booktitle   = {Proceedings of The 12th Language Resources and

```
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6504--6513},
  abstract  = {In this paper we present a multidialectal corpus
approach for building a text-to-speech voice for a new dialect in a
language with existing resources, focusing on various South American
dialects of Spanish. We first present public speech datasets for
Argentinian, Chilean, Colombian, Peruvian, Puerto Rican and
Venezuelan Spanish specifically constructed with text-to-speech
applications in mind using crowd-sourcing. We then compare the
monodialectal voices built with minimal data to a multidialectal
model built by pooling all the resources from all dialects. Our
results show that the multidialectal model outperforms the
monodialectal baseline models. We also experiment with a ``zero-
resource'' dialect scenario where we build a multidialectal voice
for a dialect while holding out target dialect recordings from the
training data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.801}
}

@InProceedings{chlbowski-ballier:2020:LREC,
  author    = {Chlébowski, Aurélie  and  Ballier, Nicolas},
  title     = {A Manually Annotated Resource for the Investigation
of Nasal Grunts},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6514--6522},
  abstract  = {This paper presents an annotation framework for nasal
grunts of the whole French CID corpus (Bertrand et al., 2008). The
acoustic components under scrutiny are justified and the annotation
guidelines are described. We carefully characterise the acoustic
cues and visual cues followed by the annotator, especially for non-
modal phonation types. The conventions followed for the annotation
of interactional and positional properties of grunts are explained.
The resulting datasets after data extraction with Praat scripts
(Boersma and Weenink, 2019) are analysed with R (R Core Team, 2017),
focusing on duration. We analyse the effect of non-modal phonation
(especially ingressive phonation) on duration and discuss a
specialisation of grunts observed in the CID for grunts with
ingressive phonation. The more general aim of this research is to
establish putative core and additive properties of grunts and a
tentative typology of grunts in spoken interactions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.802}
}

@InProceedings{martin-EtAl:2020:LREC2,
  author    = {Martin, Vincent P.  and  Rouas, Jean-Luc  and
```

```
Micoulaud Franchi, Jean-Arthur  and  Philip, Pierre},
  title     = {The Objective and Subjective Sleepiness Voice
Corpora},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6523--6531},
  abstract  = {Following patients with chronic sleep disorders
involves multiple appointments between doctors and patients which
often results in episodic follow-ups with unevenly spaced
interviews. Speech technologies and virtual doctors can help improve
this follow-up. However, there are still some challenges to
overcome: sleepiness measurements are diverse and are not always
correlated, and most past research focused on detecting nstantaneous
sleepiness levels of healthy sleep-deprived subjects. This article
presents a large database to assess the sleepiness level of highly
phenotyped patients that complain from excessive daytime sleepiness.
Based on the Multiple Sleep Latency Test, it differs from existing
databases by multiple aspects. First, it is omposed of recordings
from patients suffering from excessive daytime sleepiness instead of
sleep deprived healthy subjects. Second, it incites the subjects to
sleep contrary to existing stressing sleepiness deprivation
experimental paradigms. Third, the sleepiness level of the patients
is evaluated with different temporal granularities - long term
sleepiness and short term sleepiness - and both objective and
subjective sleepiness measures are collected. Finally, it relies on
the recordings of 94 highly phenotyped patients, allowing to unravel
the influences of different physical factors (age, sex,
weight, ... ) on voice.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.803}
}

@InProceedings{demirsahin-EtAl:2020:LREC,
  author    = {Demirsahin, Isin  and  Kjartansson, Oddur  and
Gutkin, Alexander  and  Rivera, Clara},
  title     = {Open-source Multi-speaker Corpora of the English
Accents in the British Isles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6532--6541},
  abstract  = {This paper presents a dataset of transcribed high-
quality audio of English sentences recorded by volunteers speaking
with different accents of the British Isles. The dataset is intended
for linguistic analysis as well as use for speech technologies. The
recording scripts were curated specifically for accent elicitation,
covering a variety of phonological phenomena and providing a high
phoneme coverage. The scripts include pronunciations of global
```

locations, major airlines and common personal names in different accents; and native speaker pronunciations of local words. Overlapping lines for all speakers were included for idiolect elicitation, which include the same or similar lines with other existing resources such as the CSTR VCTK corpus and the Speech Accent Archive to allow for easy comparison of personal and regional accents. The resulting corpora include over 31 hours of recordings from 120 volunteers who self-identify as native speakers of Southern England, Midlands, Northern England, Welsh, Scottish and Irish varieties of English.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.804}
}

@InProceedings{xiao-slaton-xiao:2020:LREC,
  author   = {Xiao, Yimin and Slaton, Zong-Ying and Xiao, Lu},
  title   = {TV-AfD: An Imperative-Annotated Corpus from The Big Bang Theory and Wikipedia's Articles for Deletion Discussions},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6542--6548},
  abstract = {In this study, we created an imperative corpus with speech conversations from dialogues in The Big Bang Theory and with the written comments in Wikipedia's Articles for Deletion discussions. For the TV show data, 59 episodes containing 25,076 statements are used. We manually annotated imperatives based on the annotation guideline adapted from Condoravdi and Lauer's study (2012) and used the retrieved data to assess the performance of syntax-based classification rules. For the Wikipedia AfD comments data, we first developed and leveraged a syntax-based classifier to extract 10,624 statements that may be imperative, and we manually examined the statements and then identified true positives. With this corpus, we also examined the performance of the rule-based imperative detection tool. Our result shows different outcomes for speech (dialogue) and written data. The rule-based classification performs better in the written data in precision (0.80) compared to the speech data (0.44). Also, the rule-based classification has a low-performance overall for speech data with the precision of 0.44, recall of 0.41, and f-1 measure of 0.42. This finding implies the syntax-based model may need to be adjusted for a speech dataset because imperatives in oral communication have greater syntactic varieties and are highly context-dependent.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.805}
}

@InProceedings{chen-EtAl:2020:LREC2,
  author   = {Chen, Eric and Lu, Zhiyun and Xu, Hao and Cao, Liangliang and Zhang, Yu and Fan, James},
  title   = {A Large Scale Speech Sentiment Corpus},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6549--6555},
  abstract  = {We present a multimodal corpus for sentiment analysis
based on the existing Switchboard-1 Telephone Speech Corpus released
by the Linguistic Data Consortium. This corpus extends the
Switchboard-1 Telephone Speech Corpus by adding sentiment labels
from 3 different human annotators for every transcript segment. Each
sentiment label can be one of three options: positive, negative, and
neutral. Annotators are recruited using Google Cloud's data labeling
service and the labeling task was conducted over the internet. The
corpus contains a total of 49500 labeled speech segments covering
140 hours of audio. To the best of our knowledge, this is the
largest multimodal Corpus for sentiment analysis that includes both
speech and text features.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.806}
}

@InProceedings{kachkovskaia-EtAl:2020:LREC,
  author    = {Kachkovskaia, Tatiana  and  Chukaeva, Tatiana  and
Evdokimova, Vera  and  Kholiavin, Pavel  and  Kriakina, Natalia  and
Kocharov, Daniil  and  Mamushina, Anna  and  Menshikova, Alla  and
Zimina, Svetlana},
  title     = {SibLing Corpus of Russian Dialogue Speech Designed
for Research on Speech Entrainment},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6556--6561},
  abstract  = {The paper presents a new corpus of dialogue speech
designed specifically for research in the field of speech
entrainment. Given that the degree of accommodation may depend on a
number of social factors, the corpus is designed to encompass 5
types of relations between the interlocutors: those between
siblings, close friends, strangers of the same gender, strangers of
the other gender, strangers of which one has a higher job position
and greater age. Another critical decision taken in this corpus is
that in all these social settings one speaker is kept the same. This
allows us to trace the changes in his/her speech depending on the
interlocutor. The basic set of speakers consists of 10 pairs of
same-gender siblings (including 4 pairs of identical twins) aged
23-40, and each of them was recorded in the 5 settings mentioned
above. In total we obtained 90 dialogues of 25-60 minutes each. The
speakers played a card game and a map game; they were recorded in a
soundproof studio without being able to see each other due to a non-
transparent screen between them. The corpus contains orthographic,
phonetic and prosodic annotation and is segmented into turns and
inter-pausal units.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.807}
```

```
}

@InProceedings{ramalho-freitas-rose:2020:LREC,
  author    = {Ramalho, Ana Margarida  and  Freitas, Maria João  and
Rose, Yvan},
  title     = {PhonBank and Data Sharing: Recent Developments in
European Portuguese},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6562--6570},
  abstract  = {84 104 105 115 32 112 97 112 101 114 32 112 114 101
115 101 110 116 115 32 116 104 101 32 114 101 99 101 110 116 108 121
32 112 117 98 108 105 115 104 101 100 32 82 65 77 65 76 72 79 45 69
80 32 97 110 100 32 80 72 79 78 79 68 73 83 32 99 111 114 112 111
114 97 46 32 66 111 116 104 32 105 110 99 108 117 100 101 32 69 117
114 111 112 101 97 110 32 80 111 114 116 117 103 117 101 115 101 32
112 114 111 100 117 99 116 105 111 110 32 100 97 116 97 32 102 114
111 109 32 80 111 114 116 117 103 117 101 115 101 32 99 104 105 108
100 114 101 110 32 119 105 116 104 32 116 121 112 105 99 97 108 32
40 82 65 77 65 76 72 79 45 69 80 41 32 97 110 100 32 112 114 111 116
114 97 99 116 101 100 32 40 80 72 79 78 79 68 73 83 41 32 112 104
111 110 111 108 111 103 105 99 97 108 32 100 101 118 101 108 111 112
109 101 110 116 46 32 84 104 101 32 100 97 116 97 32 105 110 32 116
104 101 32 116 119 111 32 99 111 114 112 111 114 97 32 119 101 114
101 32 99 111 108 108 101 99 116 101 100 32 117 115 105 110 103 32
116 104 101 32 112 104 111 110 111 108 111 103 105 99 97 108 32 97
115 115 101 115 115 109 101 110 116 32 116 111 111 108 32 67 76 67
80 45 69 80 44 32 100 101 118 101 108 111 112 101 100 32 105 110 32
116 104 101 32 99 111 110 116 101 120 116 32 111 102 32 116 104 101
32 67 114 111 115 115 108 105 110 103 117 105 115 116 105 99 32 67
104 105 108 100 32 80 104 111 110 111 108 111 103 121 32 80 114 111
106 101 99 116 44 32 99 111 111 114 100 105 110 97 116 101 100 32 98
121 32 66 97 114 98 97 114 97 32 66 101 114 110 104 97 114 100 116
32 97 110 100 32 74 111 101 32 83 116 101 109 98 101 114 103 101 114
32 40 85 110 105 118 101 114 115 105 116 121 32 111 102 32 66 114
105 116 105 115 104 32 67 111 108 117 109 98 105 97 32 40 85 66 67
41 44 32 67 97 110 97 100 97 41 46 32 32 66 111 116 104 32 99 111
114 112 111 114 97 32 97 114 101 32 112 97 114 116 32 111 102 32 116
104 101 32 80 104 111 110 66 97 110 107 32 80 114 111 106 101 99 116
32 40 66 114 105 97 110 32 77 97 99 87 104 105 110 110 101 121 32 40
67 97 114 110 101 103 105 101 32 77 101 108 108 111 110 44 32 85 83
65 41 32 97 110 100 32 89 118 97 110 32 82 111 115 101 32 40 77 101
109 111 114 105 97 108 32 85 110 105 118 101 114 115 105 116 121 32
111 102 32 78 101 119 102 111 117 110 100 108 97 110 100 44 32 67 97
110 97 100 97 41 44 32 119 104 105 99 104 32 105 115 32 116 104 101
32 99 104 105 108 100 32 112 104 111 110 111 108 111 103 121 32 99
111 109 112 111 110 101 110 116 32 111 102 32 84 97 108 107 66 97
110 107 44 32 99 111 111 114 100 105 110 97 116 101 100 32 98 121 32
66 114 105 97 110 32 77 97 99 87 104 105 110 110 101 121 46 32 84
104 101 32 100 97 116 97 32 97 116 32 80 104 111 110 66 97 110 107
```

32 105 115 32 101 100 105 116 101 100 32 105 110 32 80 104 111 110
32 102 111 114 109 97 116 44 32 97 32 108 97 110 103 117 97 103 101
32 116 111 111 108 32 100 101 115 105 103 110 101 100 32 97 110 100
32 98 117 105 108 116 32 98 121 32 89 118 97 110 32 82 111 115 101
32 97 110 100 32 71 114 101 103 32 72 101 100 108 117 110 100 32 40
77 101 109 111 114 105 97 108 32 85 110 105 118 101 114 115 105 116
121 32 111 102 32 78 101 119 102 111 117 110 100 108 97 110 100 41
32 97 110 100 32 119 105 100 101 108 121 32 117 115 101 100 32 98
121 32 114 101 115 101 97 114 99 104 101 114 115 32 119 111 114 107
105 110 103 32 105 110 32 116 104 101 32 102 105 101 108 100 32 111
102 32 112 104 111 110 111 108 111 103 105 99 97 108 32 97 99 113
117 105 115 105 116 105 111 110 46 32 82 65 77 65 76 72 79 45 69 80
32 99 111 110 116 97 105 110 115 32 112 114 111 100 117 99 116 105
111 110 32 100 97 116 97 32 102 114 111 109 32 56 55 32 116 121 112
105 99 97 108 108 121 32 100 101 118 101 108 111 112 105 110 103 32
99 104 105 108 100 114 101 110 44 32 97 103 101 100 32 50 59 49 49
32 116 111 32 54 59 48 52 44 32 97 108 108 32 109 111 110 111 108
105 110 103 117 97 108 115 46 32 80 72 79 78 79 68 73 83 32 105 110
99 108 117 100 101 115 32 112 114 111 100 117 99 116 105 111 110 32
100 97 116 97 32 102 114 111 109 32 50 50 32 99 104 105 108 100 114
101 110 32 100 105 97 103 110 111 115 101 100 32 119 105 116 104 32
100 105 102 102 101 114 101 110 116 32 116 121 112 101 115 32 111
102 32 115 112 101 101 99 104 32 97 110 100 32 108 97 110 103 117 97
103 101 32 100 105 115 111 114 100 101 114 115 44 32 97 108 108 32
69 80 32 109 111 110 111 108 105 110 103 117 97 108 115 44 32 97 103
101 100 32 51 59 50 32 116 111 32 49 49 44 48 53 46 32 66 111 116
104 32 99 111 114 112 111 114 97 32 97 114 101 32 111 112 101 110 32
97 99 99 101 115 115 32 108 97 110 103 117 97 103 101 32 114 101 115
111 117 114 99 101 115 32 97 110 100 32 99 111 110 116 114 105 98
117 116 101 32 116 111 32 101 110 108 97 114 103 101 32 116 104 101
32 97 109 111 117 110 116 32 111 102 32 112 114 111 100 117 99 116
105 111 110 32 100 97 116 97 32 111 110 32 116 104 101 32 97 99 113
117 105 115 105 116 105 111 110 32 111 102 32 69 117 114 111 112 101
97 110 32 80 111 114 116 117 103 117 101 115 101 32 97 118 97 105
108 97 98 108 101 32 105 110 32 80 104 111 110 66 97 110 107 46},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.808}
}

@InProceedings{saito-takamichi-saruwatari:2020:LREC,
  author   = {Saito, Yuki and Takamichi, Shinnosuke and Saruwatari, Hiroshi},
  title   = {SMASH Corpus: A Spontaneous Speech Corpus Recording Third-person Audio Commentaries on Gameplay},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6571--6577},
  abstract = {Developing a spontaneous speech corpus would be beneficial for spoken language processing and understanding. We present a speech corpus named the SMASH corpus, which includes spontaneous speech of two Japanese male commentators that made

third-person audio commentaries during the gameplay of a fighting game. Each commentator ad-libbed while watching the gameplay with various topics covering not only explanations of each moment to convey the information on the fight but also comments to entertain listeners. We made transcriptions and topic tags as annotations on the recorded commentaries with our two-step method. We first made automatic and manual transcriptions of the commentaries and then manually annotated the topic tags. This paper describes how we constructed the SMASH corpus and reports some results of the annotations.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.809}
}

@InProceedings{enomoto-den-ishimoto:2020:LREC,
  author   = {Enomoto, Mika and Den, Yasuharu and Ishimoto, Yuichi},
  title   = {A Conversation-Analytic Annotation of Turn-Taking Behavior in Japanese Multi-Party Conversation and its Preliminary Analysis},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {644--652},
  abstract = {In this study, we propose a conversation-analytic annotation scheme for turn-taking behavior in multi-party conversations. The annotation scheme is motivated by a proposal of a proper model of turn-taking incorporating various ideas developed in the literature of conversation analysis. Our annotation consists of two sets of tags: the beginning and the ending type of the utterance. Focusing on the ending-type tags, in some cases combined with the beginning-type tags, we emphasize the importance of the distinction among four selection types: i) selecting other participant as next speaker, ii) not selecting next speaker but followed by a switch of the speakership, iii) not selecting next speaker and followed by a continuation of the speakership, and iv)being inside a multi-unit turn. Based on the annotation of Japanese multi-party conversations, we analyze how syntactic and prosodic features of utterances vary across the four selection types. The results show that the above four-way distinction is essential to account for the distributions of the syntactic and prosodic features, suggesting the insufficiency of previous turn-taking models that do not consider the distinction between i) and ii) or between ii) or iii).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.81}
}

@InProceedings{fukuda-EtAl:2020:LREC,
  author   = {Fukuda, Meiko and Nishizaki, Hiromitsu and Iribe, Yurie and Nishimura, Ryota and Kitaoka, Norihide},
  title   = {Improving Speech Recognition for the Elderly: A New Corpus of Elderly Japanese Speech and Investigation of Acoustic

Modeling for Speech Recognition},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6578--6585},
  abstract = {In an aging society like Japan, a highly accurate speech recognition system is needed for use in electronic devices for the elderly, but this level of accuracy cannot be obtained using conventional speech recognition systems due to the unique features of the speech of elderly people. S-JNAS, a corpus of elderly Japanese speech, is widely used for acoustic modeling in Japan, but the average age of its speakers is 67.6 years old. Since average life expectancy in Japan is now 84.2 years, we are constructing a new speech corpus, which currently consists of the utterances of 221 speakers with an average age of 79.2, collected from four regions of Japan. In addition, we expand on our previous study (Fukuda, 2019) by further investigating the construction of acoustic models suitable for elderly speech. We create new acoustic models and train them using a combination of existing Japanese speech corpora (JNAS, S-JNAS, CSJ), with and without our ‘super-elderly’ speech data, and conduct speech recognition experiments. Our new acoustic models achieve word error rates (WER) as low as 13.38\%, exceeding the results of our previous study in which we used the CSJ acoustic model adapted for elderly speech (17.4\% WER).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.810}
}

@InProceedings{ahmed-EtAl:2020:LREC2,
  author   = {Ahmed, Shafayat and Sadeq, Nafis and Shubha, Sudipta Saha and Islam, Md. Nahidul and Adnan, Muhammad Abdullah and Islam, Mohammad Zuberul},
  title   = {Preparation of Bangla Speech Corpus from Publicly Available Audio \& Text},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6586--6592},
  abstract = {Automatic speech recognition systems require large annotated speech corpus. The manual annotation of a large corpus is very difficult. In this paper, we focus on the automatic preparation of a speech corpus for Bangladeshi Bangla. We have used publicly available Bangla audiobooks and TV news recordings as audio sources. We designed and implemented an iterative algorithm that takes as input a speech corpus and a huge amount of raw audio (without transcription) and outputs a much larger speech corpus with reasonable confidence. We have leveraged speaker diarization, gender detection, etc. to prepare the annotated corpus. We also have prepared a synthetic speech corpus for handling out-of-vocabulary

word problems in Bangla language. Our corpus is suitable for training with Kaldi. Experimental results show that the use of our corpus in addition to the Google Speech corpus (229 hours) significantly improves the performance of the ASR system.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.811}
}

@InProceedings{huang-EtAl:2020:LREC2,
  author   = {Huang, Xian and Jin, Xin and Li, Qike and Zhang, Keliang},
  title   = {On Construction of the ASR-oriented Indian English Pronunciation Dictionary},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6593--6598},
  abstract = {As a World English, a New English and a regional variety of English, Indian English (IE) has developed its own distinctive characteristics, especially phonologically, from other varieties of English. An Automatic Speech Recognition (ASR) system simply trained on British English (BE) /American English (AE) speech data and using the BE/AE pronunciation dictionary performs much worse when applied to IE. An applicable IEASR system needs spontaneous IE speech as training materials and a comprehensive, linguistically-guided IE pronunciation dictionary (IEPD) so as to achieve the effective mapping between the acoustic model and language model. This research builds a small IE spontaneous speech corpus, analyzes and summarizes the phonological variation features of IE, comes up with an IE phoneme set and complies the IEPD (including a common-English-word list, an Indian-word list, an acronym list and an affix list). Finally, two ASR systems are trained with 120 hours IE spontaneous speech data, using the IEPD we construct in this study and CMUdict separately. The two systems are tested with 50 audio clips of IE spontaneous speech. The result shows the system trained with IEPD performs better than the one trained with CMUdict with WER being 15.63\% lower on the test data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.812}
}

@InProceedings{garnerin-rossato-besacier:2020:LREC,
  author   = {Garnerin, Mahault and Rossato, Solange and Besacier, Laurent},
  title   = {Gender Representation in Open Source Speech Resources},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {6599--6605},
  abstract  = {With the rise of artificial intelligence (AI) and the
growing use of deep-learning architectures, the question of ethics,
transparency and fairness of AI systems has become a central concern
within the research community. We address transparency and fairness
in spoken language systems by proposing a study about gender
representation in speech resources available through the Open Speech
and Language Resource platform. We show that finding gender
information in open source corpora is not straightforward and that
gender balance depends on other corpus characteristics (elicited/non
elicited speech, low/high resource language, speech task targeted).
The paper ends with recommendations about metadata and gender
information for researchers in order to assure better transparency
of the speech systems built using such corpora.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.813}
}

@InProceedings{georgescu-EtAl:2020:LREC,
  author    = {Georgescu, Alexandru-Lucian  and  Cucu, Horia  and
Buzo, Andi  and  Burileanu, Corneliu},
  title     = {RSC: A Romanian Read Speech Corpus for Automatic
Speech Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6606--6612},
  abstract  = {Although many efforts have been made in the last
decade to enhance the speech and language resources for Romanian,
this language is still considered under-resourced. While for many
other languages there are large speech corpora available for
research and commercial applications, for Romanian language the
largest publicly available corpus to date comprises less than 50
hours of speech. In this context, Speech and Dialogue research group
releases Read Speech Corpus (RSC) – a Romanian speech corpus
developed in-house, comprising 100 hours of speech recordings from
164 different speakers. The paper describes the development of the
corpus and presents baseline automatic speech recognition (ASR)
results using state-of-the-art ASR technology: Kaldi speech
recognition toolkit.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.814}
}

@InProceedings{robertson-munteanu-penn:2020:LREC,
  author    = {Robertson, Sean  and  Munteanu, Cosmin  and  Penn,
Gerald},
  title     = {FAB: The French Absolute Beginner Corpus for
Pronunciation Training},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
```

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6613--6620},
  abstract  = {We introduce the French Absolute Beginner (FAB)
speech corpus. The corpus is intended for the development and study
of Computer-Assisted Pronunciation Training (CAPT) tools for
absolute beginner learners. Data were recorded during two
experiments focusing on using a CAPT system in paired role-play
tasks. The setting grants FAB three distinguishing features from
other non-native corpora: the experimental setting is ecologically
valid, closing the gap between training and deployment; it features
a label set based on teacher feedback, allowing for context-
sensitive CAPT; and data have been primarily collected from absolute
beginners, a group often ignored. Participants did not read prompts,
but instead recalled and modified dialogues that were modelled in
videos. Unable to distinguish modelled words solely from viewing
videos, speakers often uttered unintelligible or out-of-L2 words.
The corpus is split into three partitions: one from an experiment
with minimal feedback; another with explicit, word-level feedback;
and a third with supplementary read-and-record data. A subset of
words in the first partition has been labelled as more or less
native, with inter-annotator agreement reported. In the explicit
feedback partition, labels are derived from the experiment's online
feedback. The FAB corpus is scheduled to be made freely available by
the end of 2020.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.815}
}

@InProceedings{jones-EtAl:2020:LREC,
  author    = {Jones, Karen  and  Strassel, Stephanie  and  Walker,
Kevin  and  Wright, Jonathan},
  title     = {Call My Net 2: A New Resource for Speaker
Recognition},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6621--6626},
  abstract  = {We introduce the Call My Net 2 (CMN2) Corpus, a new
resource for speaker recognition featuring Tunisian Arabic
conversations between friends and family, incorporating both
traditional telephony and VoIP data. The corpus contains data from
over 400 Tunisian Arabic speakers collected via a custom-built
platform deployed in Tunis, with each speaker making 10 or more
calls each lasting up to 10 minutes. Calls include speech in various
realistic and natural acoustic settings, both noisy and non-noisy.
Speakers used a variety of handsets, including landline and mobile
devices, and made VoIP calls from tablets or computers. All calls
were subject to a series of manual and automatic quality checks,
including speech duration, audio quality, language identity and
speaker identity. The CMN2 corpus has been used in two NIST Speaker
Recognition Evaluations (SRE18 and SRE19), and the SRE test sets as
```

well as the full CMN2 corpus will be published in the Linguistic Data Consortium Catalog. We describe CMN2 corpus requirements, the telephone collection platform, and procedures for call collection. We review properties of the CMN2 dataset and discuss features of the corpus that distinguish it from prior SRE collection efforts, including some of the technical challenges encountered with collecting VoIP data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.816}
}

@InProceedings{hussain-EtAl:2020:LREC,
  author   = {Hussain, Juan and Zenkri, Oussama and Stüker, Sebastian and Waibel, Alex},
  title   = {DaCToR: A Data Collection Tool for the RELATER Project},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6627--6632},
  abstract = {Collecting domain-specific data for under-resourced languages, e.g., dialects of languages, can be very expensive, potentially financially prohibitive and taking long time. Moreover, in the case of rarely written languages, the normalization of non-canonical transcription might be another time consuming but necessary task. In order to collect domain-specific data in such circumstances in a time and cost-efficient way, collecting read data of pre-prepared texts is often a viable option. In order to collect data in the domain of psychiatric diagnosis in Arabic dialects for the project RELATER, we have prepared the data collection tool DaCToR for collecting read texts by speakers in the respective countries and districts in which the dialects are spoken. In this paper we describe our tool, its purpose within the project RELATER and the dialects which we have started to collect with the tool.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.817}
}

@InProceedings{daris-EtAl:2020:LREC2,
  author   = {Darǵis, Roberts and Paikens, Peteris and Gruzitis, Normunds and Auzina, Ilze and Akmane, Agate},
  title   = {Development and Evaluation of Speech Synthesis Corpora for Latvian},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6633--6637},
  abstract = {Text to speech (TTS) systems are necessary for all languages to ensure accessibility and availability of digital language services. Recent advances in neural speech synthesis have

eText to speech (TTS) systems are necessary for any language to ensure accessibility and availability of digital language services. Recent advances in neural speech synthesis have enabled the development of such systems with a data-driven approach that does not require significant development of language-specific tools. However, smaller languages often lack speech corpora that would be sufficient for training current neural TTS models, which require at least 30 hours of good quality audio recordings from a single speaker in a noiseless environment with matching transcriptions. Making such a corpus manually can be cost prohibitive. This paper presents an unsupervised approach to obtain a suitable corpus from unannotated recordings using automated speech recognition for transcription, as well as automated speaker segmentation and identification. The proposed method and software tools are applied and evaluated on a case study for developing a corpus suitable for Latvian speech synthesis based on Latvian public radio archive data.nabled the development of such systems with a data-driven approach that does not require much language-specific tool development. However, smaller languages often lack speech corpora that would be sufficient for training current neural TTS models, which require approximately 30 hours of good quality audio recordings from a single speaker in a noiseless environment with matching transcriptions. Making such a corpus manually can be cost prohibitive. This paper presents an unsupervised approach to obtain a suitable corpus from unannotated recordings using automated speech recognition for transcription, as well as automated speaker segmentation and identification. The proposed methods and software tools are applied and evaluated on a case study for developing a corpus suitable for Latvian speech synthesis based on Latvian public radio archive data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.818}
}

@InProceedings{nikolov-hahnloser:2020:LREC,
  author   = {Nikolov, Nikola I. and Hahnloser, Richard},
  title   = {Abstractive Document Summarization without Parallel Data},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6638--6644},
  abstract = {Abstractive summarization typically relies on large collections of paired articles and summaries. However, in many cases, parallel data is scarce and costly to obtain. We develop an abstractive summarization system that relies only on large collections of example summaries and non-matching articles. Our approach consists of an unsupervised sentence extractor that selects salient sentences to include in the final summary, as well as a sentence abstractor that is trained on pseudo-parallel and synthetic data, that paraphrases each of the extracted sentences. We perform an extensive evaluation of our method: on the CNN/DailyMail

benchmark, on which we compare our approach to fully supervised baselines, as well as on the novel task of automatically generating a press release from a scientific journal article, which is well suited for our system. We show promising performance on both tasks, without relying on any article-summary pairs.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.819}
}

@InProceedings{asao-EtAl:2020:LREC,
  author   = {Asao, Yoshihiko and Kloetzer, Julien and Mizuno, Junta and Saiki, Dai and Kadowaki, Kazuma and Torisawa, Kentaro},
  title   = {Understanding User Utterances in a Dialog System for Caregiving},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {653--661},
  abstract = {A dialog system that can monitor the health status of seniors has a huge potential for solving the labor force shortage in the caregiving industry in aging societies. As a part of efforts to create such a system, we are developing two modules that are aimed to correctly interpret user utterances: (i) a yes/no response classifier, which categorizes responses to health-related yes/no questions that the system asks; and (ii) an entailment recognizer, which detects users' voluntary mentions about their health status. To apply machine learning approaches to the development of the modules, we created large annotated datasets of 280,467 question-response pairs and 38,868 voluntary utterances. For question-response pairs, we asked annotators to avoid direct "yes" or "no" answers, so that our data could cover a wide range of possible natural language responses. The two modules were implemented by fine-tuning a BERT model, which is a recent successful neural network model. For the yes/no response classifier, the macro-average of the average precisions (APs) over all of our four categories (Yes/No/Unknown/Other) was 82.6\% (96.3\% for "yes" responses and 91.8\% for "no" responses), while for the entailment recognizer it was 89.9\%.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.82}
}

@InProceedings{antognini-faltings:2020:LREC2,
  author   = {Antognini, Diego and Faltings, Boi},
  title   = {GameWikiSum: a Novel Large Multi-Document Summarization Dataset},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

```
  pages     = {6645--6650},
  abstract  = {Today's research progress in the field of multi-
document summarization is obstructed by the small number of
available datasets. Since the acquisition of reference summaries is
costly, existing datasets contain only hundreds of samples at most,
resulting in heavy reliance on hand-crafted features or
necessitating additional, manually annotated data. The lack of large
corpora therefore hinders the development of sophisticated models.
Additionally, most publicly available multi-document summarization
corpora are in the news domain, and no analogous dataset exists in
the video game domain. In this paper, we propose GameWikiSum, a new
domain-specific dataset for multi-document summarization, which is
one hundred times larger than commonly used datasets, and in another
domain than news. Input documents consist of long professional video
game reviews as well as references of their gameplay sections in
Wikipedia pages. We analyze the proposed dataset and show that both
abstractive and extractive models can be trained on it. We release
GameWikiSum for further research: https://github.com/Diego999/
GameWikiSum.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.820}
}

@InProceedings{frefel:2020:LREC,
  author    = {Frefel, Dominik},
  title     = {Summarization Corpora of Wikipedia Articles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6651--6655},
  abstract  = {In this paper we propose a process to extract
summarization corpora from Wikipedia articles. Applied to the German
language we create a corpus of 240,000 texts. We use ROUGE scores
for the extraction and evaluation of our corpus. For this we provide
a ROUGE metric implementation adapted to the German language. The
extracted corpus is used to train three abstractive summarization
models which we compare to different baselines. The resulting
summaries sound natural and cover the input text very well. The
corpus can be downloaded at https://github.com/domfr/GeWiki.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.821}
}

@InProceedings{tauchmann-mieskes:2020:LREC,
  author    = {Tauchmann, Christopher  and  Mieskes, Margot},
  title     = {Language Agnostic Automatic Summarization
Evaluation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
```

```
  pages     = {6656--6662},
  abstract  = {So far work on automatic summarization has dealt
primarily with English data. Accordingly, evaluation methods were
primarily developed with this language in mind. In our work, we
present experiments of adapting available evaluation methods such as
ROUGE and PYRAMID to non-English data. We base our experiments on
various English and non-English homogeneous benchmark data sets as
well as a non-English heterogeneous data set. Our results indicate
that ROUGE can indeed be adapted to non-English data -- both
homogeneous and heterogeneous. Using a recent implementation of
performing an automatic PYRAMID evaluation, we also show its
adaptability to non-English data.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.822}
}

@InProceedings{ano-bojar:2020:LREC,
  author    = {Çano, Erion  and  Bojar, Ondřej},
  title     = {Two Huge Title and Keyword Generation Corpora of
Research Articles},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6663--6671},
  abstract  = {Recent developments in sequence-to-sequence learning
with neural networks have considerably improved the quality of
automatically generated text summaries and document keywords,
stipulating the need for even bigger training corpora. Metadata of
research articles are usually easy to find online and can be used to
perform research on various tasks. In this paper, we introduce two
huge datasets for text summarization (OAGSX) and keyword generation
(OAGKX) research, containing 34 million and 23 million records,
respectively. The data were retrieved from the Open Academic Graph
which is a network of research profiles and publications. We
carefully processed each record and also tried several extractive
and abstractive methods of both tasks to create performance
baselines for other researchers. We further illustrate the
performance of those methods previewing their outputs. In the near
future, we would like to apply topic modeling on the two sets to
derive subsets of research articles from more specific
disciplines.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.823}
}

@InProceedings{aburaed-saggion-chiruzzo:2020:LREC,
  author    = {AbuRa'ed, Ahmed  and  Saggion, Horacio  and
Chiruzzo, Luis},
  title     = {A Multi-level Annotated Corpus of Scientific Papers
for Scientific Document Summarization and Cross-document Relation
Discovery},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6672--6679},
  abstract  = {Related work sections or literature reviews are an
essential part of every scientific article being crucial for paper
reviewing and assessment. The automatic generation of related work
sections can be considered an instance of the multi-document
summarization problem. In order to allow the study of this specific
problem, we have developed a manually annotated, machine readable
data-set of related work sections, cited papers (e.g. references)
and sentences, together with an additional layer of papers citing
the references. We additionally present experiments on the
identification of cited sentences, using as input citation contexts.
The corpus alongside the gold standard are made available for use by
the scientific community.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.824}
}

@InProceedings{aksenov-EtAl:2020:LREC,
  author    = {Aksenov, Dmitrii  and  Moreno-Schneider, Julian  and
Bourgonje, Peter  and  Schwarzenberg, Robert  and  Hennig, Leonhard
and  Rehm, Georg},
  title     = {Abstractive Text Summarization based on Language
Model Conditioning and Locality Modeling},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6680--6689},
  abstract  = {We explore to what extent knowledge about the pre-
trained language model that is used is beneficial for the task of
abstractive summarization. To this end, we experiment with
conditioning the encoder and decoder of a Transformer-based neural
model on the BERT language model. In addition, we propose a new
method of BERT-windowing, which allows chunk-wise processing of
texts longer than the BERT window size. We also explore how locality
modeling, i.e., the explicit restriction of calculations to the
local context, can affect the summarization ability of the
Transformer. This is done by introducing 2-dimensional convolutional
self-attention into the first layers of the encoder. The results of
our models are compared to a baseline and the state-of-the-art
models on the CNN/Daily Mail dataset. We additionally train our
model on the SwissText dataset to demonstrate usability on German.
Both models outperform the baseline in ROUGE scores on two datasets
and show its superiority in a manual qualitative analysis.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.825}
}

@InProceedings{mieskes-lozamenca-kronsbein:2020:LREC,
  author    = {Mieskes, Margot  and  Loza Mencía, Eneldo  and
```

Kronsbein, Tim},
  title   = {A Data Set for the Analysis of Text Quality Dimensions in Summarization Evaluation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6690--6699},
  abstract = {Automatic evaluation of summarization focuses on developing a metric to represent the quality of the resulting text. However, text qualityis represented in a variety of dimensions ranging from grammaticality to readability and coherence. In our work, we analyze the depen-dencies between a variety of quality dimensions on automatically created multi-document summaries and which dimensions automaticevaluation metrics such as ROUGE, PEAK or JSD are able to capture. Our results indicate that variants of ROUGE are correlated tovarious quality dimensions and that some automatic summarization methods achieve higher quality summaries than others with respectto individual summary quality dimensions. Our results also indicate that differentiating between quality dimensions facilitates inspectionand fine-grained comparison of summarization methods and its characteristics. We make the data from our two summarization qualityevaluation experiments publicly available in order to facilitate the future development of specialized automatic evaluation methods.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.826}
}

@InProceedings{httasch-EtAl:2020:LREC,
  author   = {Hättasch, Benjamin and Geisler, Nadja and Meyer, Christian M. and Binnig, Carsten},
  title   = {Summarization Beyond News: The Automatically Acquired Fandom Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6700--6708},
  abstract = {Large state-of-the-art corpora for training neural networks to create abstractive summaries are mostly limited to the news genre, as it is expensive to acquire human-written summaries for other types of text at a large scale. In this paper, we present a novel automatic corpus construction approach to tackle this issue as well as three new large open-licensed summarization corpora based on our approach that can be used for training abstractive summarization models. Our constructed corpora contain fictional narratives, descriptive texts, and summaries about movies, television, and book series from different domains. All sources use a creative commons (CC) license, hence we can provide the corpora for download. In addition, we also provide a ready-to-use framework

that implements our automatic construction approach to create custom corpora with desired parameters like the length of the target summary and the number of source documents from which to create the summary. The main idea behind our automatic construction approach is to use existing large text collections (e.g., thematic wikis) and automatically classify whether the texts can be used as (query-focused) multi-document summaries and align them with potential source texts. As a final contribution, we show the usefulness of our automatic construction approach by running state-of-the-art summarizers on the corpora and through a manual evaluation with human annotators.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.827}
}

@InProceedings{demattei-EtAl:2020:LREC,
  author   = {De Mattei, Lorenzo and Cafagna, Michele and Dell'Orletta, Felice and Nissim, Malvina},
  title   = {Invisible to People but not to Machines: Evaluation of Style-aware HeadlineGeneration in Absence of Reliable Human Judgment},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6709--6717},
  abstract = {We automatically generate headlines that are expected to comply with the specific styles of two different Italian newspapers. Through a data alignment strategy and different training/testing settings, we aim at decoupling content from style and preserve the latter in generation. In order to evaluate the generated headlines' quality in terms of their specific newspaper-compliance, we devise a fine-grained evaluation strategy based on automatic classification. We observe that our models do indeed learn newspaper-specific style. Importantly, we also observe that humans aren't reliable judges for this task, since although familiar with the newspapers, they are not able to discern their specific styles even in the original human-written headlines. The utility of automatic evaluation goes therefore beyond saving the costs and hurdles of manual annotation, and deserves particular care in its design.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.828}
}

@InProceedings{tardy-EtAl:2020:LREC,
  author   = {Tardy, Paul and Janiszek, David and Estève, Yannick and Nguyen, Vincent},
  title   = {Align then Summarize: Automatic Alignment Methods for Summarization Corpus Creation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6718--6724},
  abstract  = {Summarizing texts is not a straightforward task. Before even considering text summarization, one should determine what kind of summary is expected. How much should the information be compressed? Is it relevant to reformulate or should the summary stick to the original phrasing? State-of-the-art on automatic text summarization mostly revolves around news articles. We suggest that considering a wider variety of tasks would lead to an improvement in the field, in terms of generalization and robustness. We explore meeting summarization: generating reports from automatic transcriptions. Our work consists in segmenting and aligning transcriptions with respect to reports, to get a suitable dataset for neural summarization. Using a bootstrapping approach, we provide pre-alignments that are corrected by human annotators, making a validation set against which we evaluate automatic models. This consistently reduces annotators' efforts by providing iteratively better pre-alignment and maximizes the corpus size by using annotations from our automatic alignment models. Evaluation is conducted on publicmeetings, a novel corpus of aligned public meetings. We report automatic alignment and summarization performances on this corpus and show that automatic alignment is relevant for data annotation since it leads to large improvement of almost +4 on all ROUGE scores on the summarization task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.829}
}

@InProceedings{lin-EtAl:2020:LREC2,
  author    = {Lin, Donghui  and  Otani, Masayuki  and  Okuno, Ryosuke  and  Ishida, Toru},
  title     = {Designing Multilingual Interactive Agents using Small Dialogue Corpora},
  booktitle      = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {662--667},
  abstract  = {Interactive dialogue agents like smart speakers have become more and more popular in recent years. These agents are being developed on machine learning technologies that use huge amounts of language resources. However, many entities in specialized fields are struggling to develop their own interactive agents due to a lack of language resources such as dialogue corpora, especially when the end users need interactive agents that offer multilingual support. Therefore, we aim at providing a general design framework for multilingual interactive agents in specialized domains that, it is assumed, have small or non-existent dialogue corpora. To achieve our goal, we first integrate and customize external language services for supporting multilingual functions of interactive agents. Then, we realize context-aware dialogue generation under the situation of small corpora. Third, we develop a gradual design process for
```

acquiring dialogue corpora and improving the interactive agents. We implement a multilingual interactive agent in the field of healthcare and conduct experiments to illustrate the effectiveness of the implemented agent.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.83}
}

@InProceedings{suppa-adamec:2020:LREC,
  author   = {Suppa, Marek and Adamec, Jergus},
  title   = {A Summarization Dataset of Slovak News Articles},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6725--6730},
  abstract = {As a well established NLP task, single-document summarization has seen significant interest in the past few years. However, most of the work has been done on English datasets. This is particularly noticeable in the context of evaluation where the dominant ROUGE metric assumes its input to be written in English. In this paper we aim to address both of these issues by introducing a summarization dataset of articles from a popular Slovak news site and proposing small adaptation to the ROUGE metric that make it better suited for Slovak texts. Several baselines are evaluated on the dataset, including an extractive approach based on the Multilingual version of the BERT architecture. To the best of our knowledge, the presented dataset is the first large-scale news-based summarization dataset for text written in Slovak language. It can be reproduced using the utilities available at https://github.com/NaiveNeuron/sme-sum},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.830}
}

@InProceedings{varab-schluter:2020:LREC,
  author   = {Varab, Daniel and Schluter, Natalie},
  title   = {DaNewsroom: A Large-scale Danish Summarisation Dataset},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6731--6739},
  abstract = {Dataset development for automatic summarisation systems is notoriously English-oriented. In this paper we present the first large-scale non-English language dataset specifically curated for automatic summarisation. The document-summary pairs are news articles and manually written summaries in the Danish language. There has previously been no work done to establish a Danish summarisation dataset, nor any published work on the automatic summarisation of Danish. We provide therefore the first automatic

```
summarisation dataset for the Danish language (large-scale or
otherwise). To support the comparison of future automatic
summarisation systems for Danish, we include system performance on
this dataset of strong well-established unsupervised baseline
systems, together with an oracle extractive summariser, which is the
first account of automatic summarisation system performance for
Danish. Finally, we make all code for automatically acquiring the
data freely available and make explicit how this technology can
easily be adapted in order to acquire automatic summarisation
datasets for further languages.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.831}
}

@InProceedings{lu-henchion-macnamee:2020:LREC,
  author    = {Lu, Jinghui  and  Henchion, Maeve  and  Mac Namee,
Brian},
  title     = {Diverging Divergences: Examining Variants of Jensen
Shannon Divergence for Corpus Comparison Tasks},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6740--6744},
  abstract  = {Jensen-Shannon divergence (JSD) is a distribution
similarity measurement widely used in natural language processing.
In corpus comparison tasks, where keywords are extracted to reveal
the divergence between different corpora (for example, social media
posts from proponents of different views on a political issue), two
variants of JSD have emerged in the literature. One of these uses a
weighting based on the relative sizes of the corpora being compared.
In this paper we argue that this weighting is unnecessary and, in
fact, can lead to misleading results. We recommend that this
weighted version is not used. We base this recommendation on an
analysis of the JSD variants and experiments showing how they impact
corpus comparison results as the relative sizes of the corpora being
compared change.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.832}
}

@InProceedings{bulatov-EtAl:2020:LREC,
  author    = {Bulatov, Victor  and  Alekseev, Vasiliy  and
Vorontsov, Konstantin  and  Polyudova, Darya  and  Veselova, Eugenia
and  Goncharov, Alexey  and  Egorov, Evgeny},
  title     = {TopicNet: Making Additive Regularisation for Topic
Modelling Accessible},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6745--6752},
```

```
  abstract  = {This paper introduces TopicNet, a new Python module
for topic modeling. This package, distributed under the MIT license,
focuses on bringing additive regularization topic modelling (ARTM)
to non-specialists using a general-purpose high-level language. The
module features include powerful model visualization techniques,
various training strategies, semi-automated model selection, support
for user-defined goal metrics, and a modular approach to topic model
training. Source code and documentation are available at https://
github.com/machine-intelligence-laboratory/TopicNet},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.833}
}

@InProceedings{yamaguchi-asahi-sasaki:2020:LREC,
  author    = {Yamaguchi, Kyosuke  and  Asahi, Ryoji  and  Sasaki,
Yutaka},
  title     = {SC-CoMIcs: A Superconductivity Corpus for Materials
Informatics},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6753--6760},
  abstract  = {This paper describes a novel corpus tailored for the
text mining of superconducting materials in Materials Informatics
(MI), named SuperConductivety Corpus for Materials Informatics (SC-
CoMIcs). Different from biomedical informatics, there exist very few
corpora targeting Materials Science and Engineering (MSE).
Especially, there is no sizable corpus which can be used to assist
the search of superconducting materials. A team of materials
scientists and natural language processing experts jointly designed
the annotation and constructed a corpus consisting of manually-
annotated 1,000 MSE abstracts related to superconductivity. We
conducted experiments on the corpus with a neural Named Entity
Recognition (NER) tool. The experimental results show that NER
performance over the corpus is around 77\% in terms of micro-F1,
which is comparable to human annotator agreement rates. Using the
trained NER model, we automatically annotated 9,000 abstracts and
created a term retrieval tool based on the term similarity. This
tool can find superconductivity terms relevant to a query term
within a specified Named Entity category, which demonstrates the
power of our SC-CoMIcs, efficiently providing knowledge for
Materials Informatics applications from rapidly expanding
publications.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.834}
}

@InProceedings{hagiwara-mita:2020:LREC,
  author    = {Hagiwara, Masato  and  Mita, Masato},
  title     = {GitHub Typo Corpus: A Large-Scale Multilingual
Dataset of Misspellings and Grammatical Errors},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6761--6768},
  abstract  = {The lack of large-scale datasets has been a major
hindrance to the development of NLP tasks such as spelling
correction and grammatical error correction (GEC). As a
complementary new resource for these tasks, we present the GitHub
Typo Corpus, a large-scale, multilingual dataset of misspellings and
grammatical errors along with their corrections harvested from
GitHub, a large and popular platform for hosting and sharing git
repositories. The dataset, which we have made publicly available,
contains more than 350k edits and 65M characters in more than 15
languages, making it the largest dataset of misspellings to date. We
also describe our process for filtering true typo edits based on
learned classifiers on a small annotated subset, and demonstrate
that typo edits can be identified with F1 ~ 0.9 using a very simple
classifier with only three features. The detailed analyses of the
dataset show that existing spelling correctors merely achieve an F-
measure of approx. 0.5, suggesting that the dataset serves as a new,
rich source of spelling errors that complement existing datasets.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.835}
}

@InProceedings{arase-kajiwara-chu:2020:LREC,
  author    = {Arase, Yuki  and  Kajiwara, Tomoyuki  and  Chu,
Chenhui},
  title     = {Annotation of Adverse Drug Reactions in Patients'
Weblogs},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6769--6776},
  abstract  = {Adverse drug reactions are a severe problem that
significantly degrade quality of life, or even threaten the life of
patients. Patient-generated texts available on the web have been
gaining attention as a promising source of information in this
regard. While previous studies annotated such patient-generated
content, they only reported on limited information, such as whether
a text described an adverse drug reaction or not. Further, they only
annotated short texts of a few sentences crawled from online forums
and social networking services. The dataset we present in this paper
is unique for the richness of annotated information, including
detailed descriptions of drug reactions with full context. We
crawled patient's weblog articles shared on an online patient-
networking platform and annotated the effects of drugs therein
reported. We identified spans describing drug reactions and assigned
labels for related drug names, standard codes for the symptoms of
the reactions, and types of effects. As a first dataset, we
annotated 677 drug reactions with these detailed labels based on 169
```

weblog articles by Japanese lung cancer patients. Our annotation dataset is made publicly available at our web site (https://yukiar.github.io/adr-jp/) for further research on the detection of adverse drug reactions and more broadly, on patient-generated text processing.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.836}
}

@InProceedings{rezapour-EtAl:2020:LREC,
  author   = {Rezapour, Rezvaneh and Bopp, Jutta and Fiedler, Norman and Steffen, Diana and Witt, Andreas and Diesner, Jana},
  title   = {Beyond Citations: Corpus-based Methods for Detecting the Impact of Research Outcomes on Society},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6777--6785},
  abstract = {This paper proposes, implements and evaluates a novel, corpus-based approach for identifying categories indicative of the impact of research via a deductive (top-down, from theory to data) and an inductive (bottom-up, from data to theory) approach. The resulting categorization schemes differ in substance. Research outcomes are typically assessed by using bibliometric methods, such as citation counts and patterns, or alternative metrics, such as references to research in the media. Shortcomings with these methods are their inability to identify impact of research beyond academia (bibliometrics) and considering text-based impact indicators beyond those that capture attention (altmetrics). We address these limitations by leveraging a mixed-methods approach for eliciting impact categories from experts, project personnel (deductive) and texts (inductive). Using these categories, we label a corpus of project reports per category schema, and apply supervised machine learning to infer these categories from project reports. The classification results show that we can predict deductively and inductively derived impact categories with 76.39\% and 78.81\% accuracy (F1-score), respectively. Our approach can complement solutions from bibliometrics and scientometrics for assessing the impact of research and studying the scope and types of advancements transferred from academia to society.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.837}
}

@InProceedings{fortuna-soler-wanner:2020:LREC,
  author   = {Fortuna, Paula and Soler, Juan and Wanner, Leo},
  title   = {Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},

```
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6786--6794},
  abstract  = {The field of the automatic detection of hate speech
and related concepts has raised a lot of interest in the last years.
Different datasets were annotated and classified by means of
applying different machine learning algorithms. However, few efforts
were done in order to clarify the applied categories and homogenize
different datasets. Our study takes up this demand. We analyze six
different publicly available datasets in this field with respect to
their similarity and compatibility. We conduct two different
experiments. First, we try to make the datasets compatible and
represent the dataset classes as Fast Text word vectors analyzing
the similarity between different classes in a intra and inter
dataset manner. Second, we submit the chosen datasets to the
Perspective API Toxicity classifier, achieving different
performances depending on the categories and datasets. One of the
main conclusions of these experiments is that many different
definitions are being used for equivalent concepts, which makes most
of the publicly available datasets incompatible. Grounded in our
analysis, we provide guidelines for future dataset collection and
annotation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.838}
}

@InProceedings{persing-ng:2020:LREC,
  author    = {Persing, Isaac  and  Ng, Vincent},
  title     = {Unsupervised Argumentation Mining in Student Essays},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6795--6803},
  abstract  = {State-of-the-art systems for argumentation mining are
supervised, thus relying on training data containing manually
annotated argument components and the relationships between them. To
eliminate the reliance on annotated data, we present a novel
approach to unsupervised argument mining. The key idea is to
bootstrap from a small set of argument components automatically
identified using simple heuristics in combination with reliable
contextual cues. Results on a Stab and Gurevych's corpus of 402
essays show that our unsupervised approach rivals two supervised
baselines in performance and achieves 73.5-83.7\% of the performance
of a state-of-the-art neural approach.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.839}
}

@InProceedings{rauchbauer-EtAl:2020:LREC,
  author    = {Rauchbauer, Birgit  and  Hmamouche, Youssef  and
Bigi, Brigitte  and  Prévot, Laurent  and  Ochs, Magalie  and
Chaminade, Thierry},
  title     = {Multimodal Corpus of Bidirectional Conversation of
```

Human-human and Human-robot Interaction during fMRI Scanning},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {668--675},
  abstract = {In this paper we present investigation of real-life, bi-directional conversations. We introduce the multimodal corpus derived from these natural conversations alternating between human-human and human-robot interactions. The human-robot interactions were used as a control condition for the social nature of the human-human conversations. The experimental set up consisted of conversations between the participant in a functional magnetic resonance imaging (fMRI) scanner and a human confederate or conversational robot outside the scanner room, connected via bidirectional audio and unidirectional videoconferencing (from the outside to inside the scanner). A cover story provided a framework for natural, real-life conversations about images of an advertisement campaign. During the conversations we collected a multimodal corpus for a comprehensive characterization of bi-directional conversations. In this paper we introduce this multimodal corpus which includes neural data from functional magnetic resonance imaging (fMRI), physiological data (blood flow pulse and respiration), transcribed conversational data, as well as face and eye-tracking recordings. Thus, we present a unique corpus to study human conversations including neural, physiological and behavioral data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.84}
}

@InProceedings{ocampodiaz-zhang-ng:2020:LREC,
  author   = {Ocampo Diaz, Gerardo and Zhang, Xuanming and Ng, Vincent},
  title   = {Aspect-Based Sentiment Analysis as Fine-Grained Opinion Mining},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6804--6811},
  abstract = {We show how the general fine-grained opinion mining concepts of opinion target and opinion expression are related to aspect-based sentiment analysis (ABSA) and discuss their benefits for resource creation over popular ABSA annotation schemes. Specifically, we first discuss why opinions modeled solely in terms of (entity, aspect) pairs inadequately captures the meaning of the sentiment originally expressed by authors and how opinion expressions and opinion targets can be used to avoid the loss of information. We then design a meaning-preserving annotation scheme and apply it to two popular ABSA datasets, the 2016 SemEval ABSA

Restaurant and Laptop datasets. Finally, we discuss the importance of opinion expressions and opinion targets for next-generation ABSA systems. We make our datasets publicly available for download.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.840}
}

@InProceedings{yaneva-EtAl:2020:LREC,
  author   = {Yaneva, Victoria and Ha, Le An and Baldwin, Peter and Mee, Janet},
  title   = {Predicting Item Survival for Multiple Choice Questions in a High-Stakes Medical Exam},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6812--6818},
  abstract = {One of the most resource-intensive problems in the educational testing industry relates to ensuring that newly-developed exam questions can adequately distinguish between students of high and low ability. The current practice for obtaining this information is the costly procedure of pretesting: new items are administered to test-takers and then the items that are too easy or too difficult are discarded. This paper presents the first study towards automatic prediction of an item's probability to ``survive" pretesting (item survival), focusing on human-produced MCQs for a medical exam. Survival is modeled through a number of linguistic features and embedding types, as well as features inspired by information retrieval. The approach shows promising first results for this challenging new application and for modeling the difficulty of expert-knowledge questions.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.841}
}

@InProceedings{cho-EtAl:2020:LREC,
  author   = {Cho, Won Ik and Kim, Jong In and Moon, Young Ki and Kim, Nam Soo},
  title   = {Discourse Component to Sentence (DC2S): An Efficient Human-Aided Construction of Paraphrase and Sentence Similarity Dataset},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6819--6826},
  abstract = {Assessing the similarity of sentences and detecting paraphrases is an essential task both in theory and practice, but achieving a reliable dataset requires high resource. In this paper, we propose a discourse component-based paraphrase generation for the directive utterances, which is efficient in terms of human-aided construction and content preservation. All discourse components are

```
expressed in natural language phrases, and the phrases are created
considering both speech act and topic so that the controlled
construction of the sentence similarity dataset is available. Here,
we investigate the validity of our scheme using the Korean language,
a language with diverse paraphrasing due to frequent subject drop
and scramblings. With 1,000 intent argument phrases and thus
generated 10,000 utterances, we make up a sentence similarity
dataset of practically sufficient size. It contains five sentence
pair types, including paraphrase, and displays a total volume of
about 550K. To emphasize the utility of the scheme and dataset, we
measure the similarity matching performance via conventional natural
language inference models, also suggesting the multi-lingual
extensibility.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.842}
}

@InProceedings{hayashibe:2020:LREC,
  author    = {Hayashibe, Yuta},
  title     = {Japanese Realistic Textual Entailment Corpus},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6827--6834},
  abstract  = {We perform the textual entailment (TE) corpus
construction for the Japanese Language with the following three
characteristics: First, the corpus consists of realistic sentences;
that is, all sentences are spontaneous or almost equivalent. It does
not need manual writing which causes hidden biases. Second, the
corpus contains adversarial examples. We collect challenging
examples that can not be solved by a recent pre-trained language
model. Third, the corpus contains explanations for a part of non-
entailment labels. We perform the reasoning annotation where
annotators are asked to check which tokens in hypotheses are the
reason why the relations are labeled. It makes easy to validate the
annotation and analyze system errors. The resulting corpus consists
of 48,000 realistic Japanese examples. It is the largest among
publicly available Japanese TE corpora. Additionally, it is the
first Japanese TE corpus that includes reasons for the annotation as
we know. We are planning to distribute this corpus to the NLP
community at the time of publication.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.843}
}

@InProceedings{bernardy-chatzikyriakidis:2020:LREC,
  author    = {Bernardy, Jean-Philippe  and  Chatzikyriakidis,
Stergios},
  title     = {Improving the Precision of Natural Textual Entailment
Problem Datasets},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {6835--6840},
  abstract  = {In this paper, we propose a method to modify natural
textual entailment problem datasets so that they better reflect a
more precise notion of entailment. We apply this method to a subset
of the Recognizing Textual Entailment datasets. We thus obtain a new
corpus of entailment problems, which has the following three
characteristics: 1. it is precise (does not leave out implicit
hypotheses) 2. it is based on ``real-world'' texts (i.e. most of the
premises were written for purposes other than testing textual
entailment). 3. its size is 150. Broadly, the method that we employ
is to make any missing hypotheses explicit using a crowd of experts.
We discuss the relevance of our method in improving existing NLI
datasets to be more fit for precise reasoning and we argue that this
corpus can be the basis a first step towards wide-coverage testing
of precise natural-language inference systems.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.844}
}

@InProceedings{pragst-minker-ultes:2020:LREC,
  author    = {Pragst, Louisa  and  Minker, Wolfgang  and  Ultes,
Stefan},
  title     = {Comparative Study of Sentence Embeddings for
Contextual Paraphrasing},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6841--6851},
  abstract  = {Paraphrasing is an important aspect of natural-
language generation that can produce more variety in the way
specific content is presented. Traditionally, paraphrasing has been
focused on finding different words that convey the same meaning.
However, in human-human interaction, we regularly express our
intention with phrases that are vastly different regarding both word
content and syntactic structure. Instead of exchanging only
individual words, the complete surface realisation of a sentences is
altered while still preserving its meaning and function in a
conversation. This kind of contextual paraphrasing did not yet
receive a lot of attention from the scientific community despite its
potential for the creation of more varied dialogues. In this work,
we evaluate several existing approaches to sentence encoding with
regard to their ability to capture such context-dependent
paraphrasing. To this end, we define a paraphrase classification
task that incorporates contextual paraphrases, perform dialogue act
clustering, and determine the performance of the sentence embeddings
in a sentence swapping task.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.845}
}
```

```
@InProceedings{liu-EtAl:2020:LREC,
  author    = {Liu, Tianyu  and  Xin, Zheng  and  Chang, Baobao  and
Sui, Zhifang},
  title     = {HypoNLI: Exploring the Artificial Patterns of
Hypothesis-only Bias in Natural Language Inference},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6852--6860},
  abstract  = {Many recent studies have shown that for models
trained on datasets for natural language inference (NLI), it is
possible to make correct predictions by merely looking at the
hypothesis while completely ignoring the premise. In this work, we
manage to derive adversarial examples in terms of the hypothesis-
only bias and explore eligible ways to mitigate such bias.
Specifically, we extract various phrases from the hypotheses
(artificial patterns) in the training sets, and show that they have
been strong indicators to the specific labels. We then figure out
`hard' and `easy' instances from the original test sets whose labels
are opposite to or consistent with those indications. We also set up
baselines including both pretrained models (BERT, RoBerta, XLNet)
and competitive non-pretrained models (InferSent, DAM, ESIM). Apart
from the benchmark and baselines, we also investigate two debiasing
approaches which exploit the artificial pattern modeling to mitigate
such hypothesis-only bias: down-sampling and adversarial training.
We believe those methods can be treated as competitive baselines in
NLI debiasing tasks.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.846}
}

@InProceedings{yoshinaka-kajiwara-arase:2020:LREC,
  author    = {Yoshinaka, Masato  and  Kajiwara, Tomoyuki  and
Arase, Yuki},
  title     = {SAPPHIRE: Simple Aligner for Phrasal Paraphrase with
Hierarchical Representation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6861--6867},
  abstract  = {We present SAPPHIRE, a Simple Aligner for Phrasal
Paraphrase with HIerarchical REpresentation. Monolingual phrase
alignment is a fundamental problem in natural language understanding
and also a crucial technique in various applications such as natural
language inference and semantic textual similarity assessment.
Previous methods for monolingual phrase alignment are language-
resource intensive; they require large-scale synonym/paraphrase
lexica and high-quality parsers. Different from them, SAPPHIRE
depends only on a monolingual corpus to train word embeddings.
```

Therefore, it is easily transferable to specific domains and different languages. Specifically, SAPPHIRE first obtains word alignments using pre-trained word embeddings and then expands them to phrase alignments by bilingual phrase extraction methods. To estimate the likelihood of phrase alignments, SAPPHIRE uses phrase embeddings that are hierarchically composed of word embeddings. Finally, SAPPHIRE searches for a set of consistent phrase alignments on a lattice of phrase alignment candidates. It achieves search-efficiency by constraining the lattice so that all the paths go through a phrase alignment pair with the highest alignment score. Experimental results using the standard dataset for phrase alignment evaluation show that SAPPHIRE outperforms the previous method and establishes the state-of-the-art performance.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.847}
}

@InProceedings{scherrer:2020:LREC,
  author   = {Scherrer, Yves},
  title   = {TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6868--6873},
  abstract = {This paper presents TaPaCo, a freely available paraphrase corpus for 73 languages extracted from the Tatoeba database. Tatoeba is a crowdsourcing project mainly geared towards language learners. Its aim is to provide example sentences and translations for particular linguistic constructions and words. The paraphrase corpus is created by populating a graph with Tatoeba sentences and equivalence links between sentences "meaning the same thing". This graph is then traversed to extract sets of paraphrases. Several language-independent filters and pruning steps are applied to remove uninteresting sentences. A manual evaluation performed on three languages shows that between half and three quarters of inferred paraphrases are correct and that most remaining ones are either correct but trivial, or near-paraphrases that neutralize a morphological distinction. The corpus contains a total of 1.9 million sentences, with 200 - 250 000 sentences per language. It covers a range of languages for which, to our knowledge, no other paraphrase dataset exists. The dataset is available at https://doi.org/10.5281/zenodo.3707949.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.848}
}

@InProceedings{sathe-EtAl:2020:LREC,
  author   = {Sathe, Aalok and Ather, Salar and Le, Tuan Manh and Perry, Nathan and Park, Joonsuk},
  title   = {Automated Fact-Checking of Claims from Wikipedia},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6874--6882},
  abstract  = {Automated fact checking is becoming increasingly
vital as both truthful and fallacious information accumulate online.
Research on fact checking has benefited from large-scale datasets
such as FEVER and SNLI. However, such datasets suffer from limited
applicability due to the synthetic nature of claims and/or evidence
written by annotators that differ from real claims and evidence on
the internet. To this end, we present WikiFactCheck-English, a
dataset of 124k+ triples consisting of a claim, context and an
evidence document extracted from English Wikipedia articles and
citations, as well as 34k+ manually written claims that are refuted
by the evidence documents. This is the largest fact checking dataset
consisting of real claims and evidence to date; it will allow the
development of fact checking systems that can better process claims
and evidence in the real world. We also show that for the NLI
subtask, a logistic regression system trained using existing and
novel features achieves peak accuracy of 68\%, providing a
competitive baseline for future work. Also, a decomposable attention
model trained on SNLI significantly underperforms the models trained
on this dataset, suggesting that models trained on manually
generated data may not be sufficiently generalizable or suitable for
fact checking real-world claims.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.849}
}

@InProceedings{ochs-EtAl:2020:LREC,
  author    = {Ochs, Magalie  and  Bertrand, Roxane  and  Goujon,
Aurélie  and  Bolger, Deirdre  and  Dubarry, Anne-Sophie  and
Blache, Philippe},
  title     = {The Brain-IHM Dataset: a New Resource for Studying
the Brain Basis of Human-Human and Human-Machine Conversations},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {676--683},
  abstract  = {This paper presents an original dataset of controlled
interactions, focusing on the study of feedback items. It consists
on recordings of different conversations between a doctor and a
patient, played by actors. In this corpus, the patient is mainly a
listener and produces different feedbacks, some of them being
(voluntary) incongruent. Moreover, these conversations have been re-
synthesized in a virtual reality context, in which the patient is
played by an artificial agent. The final corpus is made of different
movies of human-human conversations plus the same conversations
replayed in a human-machine context, resulting in the first human-
human/human-machine parallel corpus. The corpus is then enriched
with different multimodal annotations at the verbal and non-verbal
```

```
levels. Moreover, and this is the first dataset of this type, we
have designed an experiment during which different participants had
to watch the movies and give an evaluation of the interaction.
During this task, we recorded participant’s brain signal. The Brain-
IHM dataset is then conceived with a triple purpose: 1/ studying
feedbacks by comparing congruent vs. incongruent feedbacks 2/
comparing human-human and human-machine production of feedbacks 3/
studying the brain basis of feedback perception.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.85}
}

@InProceedings{paulpanenghat-EtAl:2020:LREC,
  author    = {Paul Panenghat, Mithun  and  Suntwal, Sandeep  and
Rafique, Faiz  and  Sharp, Rebecca  and  Surdeanu, Mihai},
  title     = {Towards the Necessity for Debiasing Natural Language
Inference Datasets},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6883--6888},
  abstract  = {Modeling natural language inference is a challenging
task. With large annotated data sets available it has now become
feasible to train complex neural network based inference methods
which achieve state of the art performance. However, it has been
shown that these models also learn from the subtle biases inherent
in these datasets \cite{gururangan2018annotation}. In this work we
explore two techniques for delexicalization that modify the datasets
in such a way that we can control the importance that neural-network
based methods place on lexical entities. We demonstrate that the
proposed methods not only maintain the performance in-domain but
also improve performance in some out-of-domain settings. For
example, when using the delexicalized version of the FEVER dataset,
the in-domain performance of a state of the art neural network
method dropped only by 1.12\% while its out-of-domain performance on
the FNC dataset improved by 4.63\%. We release the delexicalized
versions of three common datasets used in natural language
inference. These datasets are delexicalized using two methods: one
which replaces the lexical entities in an overlap-aware manner, and
a second, which additionally incorporates semantic lifting of nouns
and verbs to their WordNet hypernym synsets},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.850}
}

@InProceedings{cardon-grabar:2020:LREC,
  author    = {Cardon, Rémi  and  Grabar, Natalia},
  title     = {A French Corpus for Semantic Similarity},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {6889--6894},
  abstract  = {Semantic similarity is an area of Natural Language
Processing that is useful for several downstream applications, such
as machine translation, natural language generation, information
retrieval, or question answering. The task consists in assessing the
extent to which two sentences express or do not express the same
meaning. To do so, corpora with graded pairs of sentences are
required. The grade is positioned on a given scale, usually going
from 0 (completely unrelated) to 5 (equivalent semantics). In this
work, we introduce such a corpus for French, the first that we know
of. It is comprised of 1,010 sentence pairs with grades from five
annotators. We describe the annotation process, analyse these data,
and perform a few experiments for the automatic grading of semantic
similarity.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.851}
}

@InProceedings{watarai-tsuchiya:2020:LREC,
  author    = {Watarai, Takuto  and  Tsuchiya, Masatoshi},
  title     = {Developing Dataset of Japanese Slot Filling Quizzes
Designed for Evaluation of Machine Reading Comprehension},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6895--6901},
  abstract  = {This paper describes our developing dataset of
Japanese slot filling quizzes designed for evaluation of machine
reading comprehension. The dataset consists of quizzes automatically
generated from Aozora Bunko, and each quiz is defined as a 4-tuple:
a context passage, a query holding a slot, an answer character and a
set of possible answer characters. The query is generated from the
original sentence, which appears immediately after the context
passage on the target book, by replacing the answer character into
the slot. The set of possible answer characters consists of the
answer character and the other characters who appear in the context
passage. Because the context passage and the query shares the same
context, a machine which precisely understand the context may select
the correct answer from the set of possible answer characters. The
unique point of our approach is that we focus on characters of
target books as slots to generate queries from original sentences,
because they play important roles in narrative texts and precise
understanding their relationship is necessary for reading
comprehension. To extract characters from target books, manually
created dictionaries of characters are employed because some
characters appear as common nouns not as named entities.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.852}
}

@InProceedings{jimnezzafra-EtAl:2020:LREC,
  author    = {Jiménez-Zafra, Salud María  and  Morante, Roser  and
```

Blanco, Eduardo and Martín Valdivia, María Teresa and Ureña López, L. Alfonso},
  title   = {Detecting Negation Cues and Scopes in Spanish},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6902--6911},
  abstract = {In this work we address the processing of negation in Spanish. We first present a machine learning system that processes negation in Spanish. Specifically, we focus on two tasks: i) negation cue detection and ii) scope identification. The corpus used in the experimental framework is the SFU Corpus. The results for cue detection outperform state-of-the-art results, whereas for scope detection this is the first system that performs the task for Spanish. Moreover, we provide a qualitative error analysis aimed at understanding the limitations of the system and showing which negation cues and scopes are straightforward to predict automatically, and which ones are challenging.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.853}
}

@InProceedings{putra-EtAl:2020:LREC,
  author   = {Putra, Jan Wira Gotama and Teufel, Simone and Matsumura, Kana and Tokunaga, Takenobu},
  title   = {TIARA: A Tool for Annotating Discourse Relations and Sentence Reordering},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6912--6920},
  abstract = {This paper introduces TIARA, a new publicly available web-based annotation tool for discourse relations and sentence reordering. Annotation tasks such as these, which are based on relations between large textual objects, are inherently hard to visualise without either cluttering the display and/or confusing the annotators. TIARA deals with the visual complexity during the annotation process by systematically simplifying the layout, and by offering interactive visualisation, including coloured links, indentation, and dual-view. TIARA's text view allows annotators to focus on the analysis of logical sequencing between sentences. A separate tree view allows them to review their analysis in terms of the overall discourse structure. The dual-view gives it an edge over other discourse annotation tools and makes it particularly attractive as an educational tool (e.g., for teaching students how to argue more effectively). As it is based on standard web technologies and can be easily customised to other annotation schemes, it can be easily used by anybody. Apart from the project it was originally designed for, in which hundreds of texts were

annotated by three annotators, TIARA has already been adopted by a second discourse annotation study, which uses it in the teaching of argumentation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.854}
}

@InProceedings{elhaj-EtAl:2020:LREC,
  author   = {El-Haj, Mahmoud and Rutherford, Nathan and Coole, Matthew and Ezeani, Ignatius and Prentice, Sheryl and Ide, Nancy and Knight, Jo and Piao, Scott and Mariani, John and Rayson, Paul and Suderman, Keith},
  title   = {Infrastructure for Semantic Annotation in the Genomics Domain},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6921--6929},
  abstract = {We describe a novel super-infrastructure for biomedical text mining which incorporates an end-to-end pipeline for the collection, annotation, storage, retrieval and analysis of biomedical and life sciences literature, combining NLP and corpus linguistics methods. The infrastructure permits extreme-scale research on the open access PubMed Central archive. It combines an updatable Gene Ontology Semantic Tagger (GOST) for entity identification and semantic markup in the literature, with a NLP pipeline scheduler (Buster) to collect and process the corpus, and a bespoke columnar corpus database (LexiDB) for indexing. The corpus database is distributed to permit fast indexing, and provides a simple web front-end with corpus linguistics methods for sub-corpus comparison and retrieval. GOST is also connected as a service in the Language Application (LAPPS) Grid, in which context it is interoperable with other NLP tools and data in the Grid and can be combined with them in more complex workflows. In a literature based discovery setting, we have created an annotated corpus of 9,776 papers with 5,481,543 words.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.855}
}

@InProceedings{shah-demelo:2020:LREC,
  author   = {Shah, Kshitij and de Melo, Gerard},
  title   = {Correcting the Autocorrect: Context-Aware Typographical Error Correction via Training Data Augmentation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6930--6936},
  abstract = {In this paper, we explore the artificial generation of typographical errors based on real-world statistics. We first

draw on a small set of annotated data to compute spelling error statistics. These are then invoked to introduce errors into substantially larger corpora. The generation methodology allows us to generate particularly challenging errors that require context-aware error detection. We use it to create a set of English language error detection and correction datasets. Finally, we examine the effectiveness of machine learning models for detecting and correcting errors based on this data.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.856}
}

@InProceedings{downs-EtAl:2020:LREC,
  author   = {Downs, Brody and Anuyah, Oghenemaro and Shukla, Aprajita and Fails, Jerry Alan and Pera, Sole and Wright, Katherine and Kennington, Casey},
  title   = {KidSpell: A Child-Oriented, Rule-Based, Phonetic Spellchecker},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6937--6946},
  abstract = {For help with their spelling errors, children often turn to spellcheckers integrated in software applications like word processors and search engines. However, existing spellcheckers are usually tuned to the needs of traditional users (i.e., adults) and generally prove unsatisfactory for children. Motivated by this issue, we introduce KidSpell, an English spellchecker oriented to the spelling needs of children. KidSpell applies (i) an encoding strategy for mapping both misspelled words and spelling suggestions to their phonetic keys and (ii) a selection process that prioritizes candidate spelling suggestions that closely align with the misspelled word based on their respective keys. To assess the effectiveness of, we compare the model's performance against several popular, mainstream spellcheckers in a number of offline experiments using existing and novel datasets. The results of these experiments show that KidSpell outperforms existing spellcheckers, as it accurately prioritizes relevant spelling corrections when handling misspellings generated by children in both essay writing and online search tasks. As a byproduct of our study, we create two new datasets comprised of spelling errors generated by children from hand-written essays and web search inquiries, which we make available to the research community.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.857}
}

@InProceedings{seeha-EtAl:2020:LREC,
  author   = {Seeha, Suteera and Bilan, Ivan and Mamani Sanchez, Liliana and Huber, Johannes and Matuschek, Michael and Schütze, Hinrich},
  title   = {ThaiLMCut: Unsupervised Pretraining for Thai Word Segmentation},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6947--6957},
  abstract  = {We propose ThaiLMCut, a semi-supervised approach for
Thai word segmentation which utilizes a bi-directional character
language model (LM) as a way to leverage useful linguistic knowledge
from unlabeled data. After the language model is trained on
substantial unlabeled corpora, the weights of its embedding and
recurrent layers are transferred to a supervised word segmentation
model which continues fine-tuning them on a word segmentation task.
Our experimental results demonstrate that applying the LM always
leads to a performance gain, especially when the amount of labeled
data is small. In such cases, the F1 Score increased by up to 2.02\
%. Even on abig labeled dataset, a small improvement gain can still
be obtained. The approach has also shown to be very beneficial for
out-of-domain settings with a gain in F1 Score of up to 3.13\%.
Finally, we show that ThaiLMCut can outperform other open source
state-of-the-art models achieving an F1 Score of 98.78\% on the
standard benchmark, InterBEST2009.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.858}
}

@InProceedings{alatrash-EtAl:2020:LREC,
  author    = {Alatrash, Reem  and  Schlechtweg, Dominik  and  Kuhn,
Jonas  and  Schulte im Walde, Sabine},
  title     = {CCOHA: Clean Corpus of Historical American English},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6958--6966},
  abstract  = {Modelling language change is an increasingly
important area of interest within the fields of sociolinguistics and
historical linguistics. In recent years, there has been a growing
number of publications whose main concern is studying changes that
have occurred within the past centuries. The Corpus of Historical
American English (COHA) is one of the most commonly used large
corpora in diachronic studies in English. This paper describes
methods applied to the downloadable version of the COHA corpus in
order to overcome its main limitations, such as inconsistent lemmas
and malformed tokens, without compromising its qualitative and
distributional properties. The resulting corpus CCOHA contains a
larger number of cleaned word tokens which can offer better insights
into language change and allow for a larger variety of tasks to be
performed.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.859}
}
```

@InProceedings{bonial-EtAl:2020:LREC,
  author   = {Bonial, Claire and Donatelli, Lucia and Abrams, Mitchell and Lukin, Stephanie M. and Tratz, Stephen and Marge, Matthew and Artstein, Ron and Traum, David and Voss, Clare},
  title   = {Dialogue-AMR: Abstract Meaning Representation for Dialogue},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {684--695},
  abstract = {This paper describes a schema that enriches Abstract Meaning Representation (AMR) in order to provide a semantic representation for facilitating Natural Language Understanding (NLU) in dialogue systems. AMR offers a valuable level of abstraction of the propositional content of an utterance; however, it does not capture the illocutionary force or speaker's intended contribution in the broader dialogue context (e.g., make a request or ask a question), nor does it capture tense or aspect. We explore dialogue in the domain of human-robot interaction, where a conversational robot is engaged in search and navigation tasks with a human partner. To address the limitations of standard AMR, we develop an inventory of speech acts suitable for our domain, and present "Dialogue-AMR", an enhanced AMR that represents not only the content of an utterance, but the illocutionary force behind it, as well as tense and aspect. To showcase the coverage of the schema, we use both manual and automatic methods to construct the "DialAMR" corpus—a corpus of human-robot dialogue annotated with standard AMR and our enriched Dialogue-AMR schema. Our automated methods can be used to incorporate AMR into a larger NLU pipeline supporting human-robot dialogue.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.86}
}

@InProceedings{zouhar-bojar:2020:LREC,
  author   = {Zouhar, Vilém and Bojar, Ondřej},
  title   = {Outbound Translation User Interface Ptakopět: A Pilot Study},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6967--6975},
  abstract = {It is not uncommon for Internet users to have to produce a text in a foreign language they have very little knowledge of and are unable to verify the translation quality. We call the task "outbound translation" and explore it by introducing an open-source modular system Ptakopět. Its main purpose is to inspect human interaction with MT systems enhanced with additional subsystems, such as backward translation and quality estimation. We follow up

with an experiment on (Czech) human annotators tasked to produce questions in a language they do not speak (German), with the help of Ptakopět. We focus on three real-world use cases (communication with IT support, describing administrative issues and asking encyclopedic questions) from which we gain insight into different strategies users take when faced with outbound translation tasks. Round trip translation is known to be unreliable for evaluating MT systems but our experimental evaluation documents that it works very well for users, at least on MT systems of mid-range quality.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.860}
}

@InProceedings{titeux-EtAl:2020:LREC,
  author   = {Titeux, Hadrien and Riad, Rachid and Cao, Xuan-Nga and Hamilakis, Nicolas and Madden, Kris and Cristia, Alejandrina and Bachoud-Lévi, Anne-Catherine and Dupoux, Emmanuel},
  title   = {Seshat: a Tool for Managing and Verifying Annotation Campaigns of Audio Data},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6976--6982},
  abstract = {We introduce Seshat, a new, simple and open-source software to efficiently manage annotations of speech corpora. The Seshat software allows users to easily customise and manage annotations of large audio corpora while ensuring compliance with the formatting and naming conventions of the annotated output files. In addition, it includes procedures for checking the content of annotations following specific rules that can be implemented in personalised parsers. Finally, we propose a double-annotation mode, for which Seshat computes automatically an associated inter-annotator agreement with the gamma measure taking into account the categorisation and segmentation discrepancies.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.861}
}

@InProceedings{costello-EtAl:2020:LREC,
  author   = {Costello, Cash and Anderson, Shelby and Bishop, Caitlyn and Mayfield, James and McNamee, Paul},
  title   = {Dragonfly: Advances in Non-Speaker Annotation for Low Resource Languages},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6983--6987},
  abstract = {Dragonfly is an open source software tool that supports annotation of text in a low resource language by non-

```
speakers of the language. Using semantic and contextual information,
non-speakers of a language familiar with the Latin script can
produce high quality named entity annotations to support
construction of a name tagger. We describe a procedure for
annotating low resource languages using Dragonfly that others can
use, which we developed based on our experience annotating data in
more than ten languages. We also present performance comparisons
between models trained on native speaker and non-speaker
annotations.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.862}
}

@InProceedings{koeva-obreshkov-yalamov:2020:LREC,
  author    = {Koeva, Svetla  and  Obreshkov, Nikola  and  Yalamov,
Martin},
  title     = {Natural Language Processing Pipeline to Annotate
Bulgarian Legislative Documents},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6988--6994},
  abstract  = {The paper presents the Bulgarian MARCELL corpus, part
of a recently developed multilingual corpus representing the
national legislation in seven European countries and the NLP
pipeline that turns the web crawled data into structured,
linguistically annotated dataset. The Bulgarian data is web crawled,
extracted from the original HTML format, filtered by document type,
tokenised, sentence split, tagged and lemmatised with a fine-grained
version of the Bulgarian Language Processing Chain, dependency
parsed with NLP- Cube, annotated with named entities (persons,
locations, organisations and others), noun phrases, IATE terms and
EuroVoc descriptors. An orchestrator process has been developed to
control the NLP pipeline performing an end-to-end data processing
and annotation starting from the documents identification and ending
in the generation of statistical reports. The Bulgarian MARCELL
corpus consists of 25,283 documents (at the beginning of November
2019), which are classified into eleven types.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.863}
}

@InProceedings{forkel-list:2020:LREC,
  author    = {Forkel, Robert  and  List, Johann-Mattis},
  title     = {CLDFBench: Give Your Cross-Linguistic Data a Lift},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {6995--7002},
  abstract  = {While the amount of cross-linguistic data is
```

```
constantly increasing, most datasets produced today and in the past
cannot be considered FAIR (findable, accessible, interoperable, and
reproducible). To remedy this and to increase the comparability of
cross-linguistic resources, it is not enough to set up standards and
best practices for data to be collected in the future. We also need
consistent workflows for the ``retro-standardization'' of data that
has been published during the past decades and centuries. With the
Cross-Linguistic Data Formats initiative, first standards for cross-
linguistic data have been presented and successfully tested. So far,
however, CLDF creation was hampered by the fact that it required a
considerable degree of computational proficiency. With cldfbench, we
introduce a framework for the retro-standardization of legacy data
and the curation of new datasets that drastically simplifies the
creation of CLDF by providing a consistent, reproducible workflow
that rigorously supports version control and long term archiving of
research data and code. The framework is distributed in form of a
Python package along with usage information and examples for best
practice. This study introduces the new framework and illustrates
how it can be applied by showing how a resource containing
structural and lexical data for Sinitic languages can be efficiently
retro-standardized and analyzed.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.864}
}

@InProceedings{machlek:2020:LREC1,
  author    = {Machálek, Tomáš},
  title     = {KonText: Advanced and Flexible Corpus Query
Interface},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7003--7008},
  abstract  = {We present an advanced, highly customizable corpus
query interface KonText built on top of core libraries of the open-
source corpus search engine NoSketch Engine (NoSkE). The aim is to
overcome some limitations of the original NoSkE user interface and
provide integration capabilities allowing connection of the basic
search service with other language resources (LRs). The introduced
features are based on long-term feedback given by the users and
researchers of the Czech National Corpus (CNC) along with other LRs
providers running KonText as a part of their services. KonText is a
fully operational and mature software deployed at the CNC since 2014
that currently handles thousands user queries per day.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.865}
}

@InProceedings{machlek:2020:LREC2,
  author    = {Machálek, Tomáš},
  title     = {Word at a Glance: Modular Word Profile Aggregator},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7009--7014},
  abstract  = {Word at a Glance (WaG) is a word profile aggregator
that provides means for exploring individual words, their comparison
and translation, based on existing language resources and related
software services. It is designed as a building kit-like application
that fetches data from different sources and compiles them into a
single, comprehensible and structured web page. WaG can be easily
configured to support many tasks, but in general, it is intended to
be used not only by language experts but also the general public.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.866}
}

@InProceedings{kupietz-diewald-margaretha:2020:LREC,
  author    = {Kupietz, Marc  and  Diewald, Nils  and  Margaretha,
Eliza},
  title     = {RKorAPClient: An R Package for Accessing the German
Reference Corpus DeReKo via KorAP},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7015--7021},
  abstract  = {Making corpora accessible and usable for linguistic
research is a huge challenge in view of (too) big data, legal issues
and a rapidly evolving methodology. This does not only affect the
design of user-friendly graphical interfaces to corpus analysis
tools, but also the availability of programming interfaces
supporting access to the functionality of these tools from various
analysis and development environments. RKorAPClient is a new
research tool in the form of an R package that interacts with the
Web API of the corpus analysis platform KorAP, which provides access
to large annotated corpora, including the German reference corpus
DeReKo with 45 billion tokens.In addition to optionally
authenticated KorAP API access, RKorAPClient provides further
processing and visualization features to simplify common corpus
analysis tasks. This paper introduces the basic functionality of
RKorAPClient and exemplifies various analysis tasks based on DeReKo,
that are bundled within the R package and can serve as a basic
framework for advanced analysis and visualization approaches.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.867}
}

@InProceedings{obeid-EtAl:2020:LREC,
  author    = {Obeid, Ossama  and  Zalmout, Nasser  and  Khalifa,
Salam  and  Taji, Dima  and  Oudah, Mai  and  Alhafni, Bashar  and
Inoue, Go  and  Eryani, Fadhl  and  Erdmann, Alexander  and  Habash,
Nizar},
  title     = {CAMeL Tools: An Open Source Python Toolkit for Arabic
```

Natural Language Processing},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7022--7032},
  abstract = {We present CAMeL Tools, a collection of open-source tools for Arabic natural language processing in Python. CAMeL Tools currently provides utilities for pre-processing, morphological modeling, Dialect Identification, Named Entity Recognition and Sentiment Analysis. In this paper, we describe the design of CAMeL Tools and the functionalities it provides.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.868}
}

@InProceedings{oliver-mikeleni:2020:LREC,
  author   = {Oliver, Antoni and Mikelenić, Bojana},
  title   = {ReSiPC: a Tool for Complex Searches in Parallel Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7033--7037},
  abstract = {In this paper, a tool specifically designed to allow for complex searches in large parallel corpora is presented. The formalism for the queries is very powerful as it uses standard regular expressions that allow for complex queries combining word forms, lemmata and POS-tags. As queries are performed over POS-tags, at least one of the languages in the parallel corpus should be POS-tagged. Searches can be performed in one of the languages or in both languages at the same time. The program is able to POS-tag the corpora using the Freeling analyzer through its Python API. ReSiPC is developed in Python version 3 and it is distributed under a free license (GNU GPL). The tool can be used to provide data for contrastive linguistics research and an example of use in a Spanish-Croatian parallel corpus is presented. ReSiPC is designed for queries in POS-tagged corpora, but it can be easily adapted for querying corpora containing other kinds of information.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.869}
}

@InProceedings{ito-EtAl:2020:LREC,
  author   = {Ito, Koichiro and Murata, Masaki and Ohno, Tomohiro and Matsubara, Shigeki},
  title   = {Relation between Degree of Empathy for Narrative Speech and Type of Responsive Utterance in Attentive Listening},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {696--701},
  abstract  = {Nowadays, spoken dialogue agents such as
communication robots and smart speakers listen to narratives of
humans. In order for such an agent to be recognized as a listener of
narratives and convey the attitude of attentive listening, it is
necessary to generate responsive utterances. Moreover, responsive
utterances can express empathy to narratives and showing an
appropriate degree of empathy to narratives is significant for
enhancing speaker's motivation. The degree of empathy shown by
responsive utterances is thought to depend on their type. However,
the relation between responsive utterances and degrees of the
empathy has not been explored yet. This paper describes the
classification of responsive utterances based on the degree of
empathy in order to explain that relation. In this research,
responsive utterances are classified into five levels based on the
effect of utterances and literature on attentive listening.
Quantitative evaluations using 37,995 responsive utterances showed
the appropriateness of the proposed classification.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.87}
}

@InProceedings{limalopez-EtAl:2020:LREC2,
  author    = {Lima Lopez, Salvador  and  Perez, Naiara  and
García-Sardiña, Laura  and  Cuadros, Montse},
  title     = {HitzalMed: Anonymisation of Clinical Text in
Spanish},
  booktitle     = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month         = {May},
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {7038--7043},
  abstract  = {HitzalMed is a web-framed tool that performs
automatic detection of sensitive information in clinical texts using
machine learning algorithms reported to be competitive for the task.
Moreover, once sensitive information is detected, different
anonymisation techniques are implemented that are configurable by
the user —for instance, substitution, where sensitive items are
replaced by same category text in an effort to generate a new
document that looks as natural as the original one. The tool is able
to get data from different document formats and outputs downloadable
anonymised data. This paper presents the anonymisation and
substitution technology and the demonstrator which is publicly
available at https://snlt.vicomtech.org/hitzalmed.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.870}
}

@InProceedings{indig-sass-mittelholcz:2020:LREC,
  author    = {Indig, Balázs  and  Sass, Bálint  and  Mittelholcz,
Iván},
```

```
  title     = {The xtsv Framework and the Twelve Virtues of
Pipelines},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7044--7052},
  abstract  = {We present xtsv, an abstract framework for building
NLP pipelines. It covers several kinds of functionalities which can
be implemented at an abstract level. We survey these features and
argue that all are desired in a modern pipeline. The framework has a
simple yet powerful internal communication format which is
essentially tsv (tab separated values) with header plus some
additional features. We put emphasis on the capabilities of the
presented framework, for example its ability to allow new modules to
be easily integrated or replaced, or the variety of its usage
options. When a module is put into xtsv, all functionalities of the
system are immediately available for that module, and the module can
be be a part of an xtsv pipeline. The design also allows convenient
investigation and manual correction of the data flow from one module
to another. We demonstrate the power of our framework with a
successful application: a concrete NLP pipeline for Hungarian called
e-magyar text processing system (emtsv) which integrates Hungarian
NLP tools in xtsv. All the advantages of the pipeline come from the
inherent properties of the xtsv framework.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.871}
}

@InProceedings{daudert:2020:LREC,
  author    = {Daudert, Tobias},
  title     = {A Web-based Collaborative Annotation and
Consolidation Tool},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7053--7059},
  abstract  = {Annotation tools are a valuable asset for the
construction of labelled textual datasets. However, they tend to
have a rigid structure, closed back-end and front-end, and are built
in a non-user-friendly way. These downfalls difficult their use in
annotation tasks requiring varied text formats, prevent researchers
to optimise the tool to the annotation task, and impede people with
little programming knowledge to easily modify the tool rendering it
unusable for a large cohort. Targeting these needs, we present a
web-based collaborative annotation and consolidation tool (AWOCATo),
capable of supporting varied textual formats. AWOCATo is based on
three pillars: (1) Simplicity, built with a modular architecture
employing easy to use technologies; (2) Flexibility, the JSON
configuration file allows an easy adaption to the annotation task;
```

(3) Customizability, parameters such as labels, colours, or consolidation features can be easily customized. These features allow AWOCATo to support a range of tasks and domains, filling the gap left by the absence of annotation tools that can be used by people with and without programming knowledge, including those who wish to easily adapt a tool to less common tasks. AWOCATo is available for download at https://github.com/TDaudert/AWOCATo.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.872}
}

@InProceedings{larson-guldan-leach:2020:LREC,
  author   = {Larson, Stefan and Guldan, Eric and Leach, Kevin},
  title   = {Data Query Language and Corpus Tools for Slot-Filling and Intent Classification Data},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7060--7068},
  abstract = {Typical machine learning approaches to developing task-oriented dialog systems require the collection and management of large amounts of training data, especially for the tasks of intent classification and slot-filling. Managing this data can be cumbersome without dedicated tools to help the dialog system designer understand the nature of the data. This paper presents a toolkit for analyzing slot-filling and intent classification corpora. We present a toolkit that includes (1) a new lightweight and readable data and file format for intent classification and slot-filling corpora, (2) a new query language for searching intent classification and slot-filling corpora, and (3) tools for understanding the structure and makeup for such corpora. We apply our toolkit to several well-known NLU datasets, and demonstrate that our toolkit can be used to uncover interesting and surprising insights. By releasing our toolkit to the research community, we hope to enable others to develop more robust and intelligent slot-filling and intent classification models.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.873}
}

@InProceedings{krishna-EtAl:2020:LREC,
  author   = {Krishna, Amrith and Vidhyut, Shiv and Chawla, Dilpreet and Sambhavi, Sruti and Goyal, Pawan},
  title   = {SHR++: An Interface for Morpho-syntactic Annotation of Sanskrit Corpora},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7069--7076},

```
  abstract  = {We propose a web-based annotation framework, SHR++,
for morpho-syntactic annotation of corpora in Sanskrit. SHR++ is
designed to generate annotations for the word-segmentation,
morphological parsing and dependency analysis tasks in Sanskrit. It
incorporates analyses and predictions from various tools designed
for processing texts in Sanskrit, and utilise them to ease the
cognitive load of the human annotators. Specifically, SHR++ uses
Sanskrit Heritage Reader, a lexicon driven shallow parser for
enumerating all the phonetically and lexically valid word splits
along with their morphological analyses for a given string. This
would help the annotators in choosing the solutions, rather than
performing the segmentations by themselves. Further, predictions
from a word segmentation tool are added as suggestions that can aid
the human annotators in their decision making. Our evaluation shows
that enabling this segmentation suggestion component reduces the
annotation time by 20.15 \%. SHR++ can be accessed online at http://
vidhyut97.pythonanywhere.com/ and the codebase, for the independent
deployment of the system elsewhere, is hosted at https://github.com/
iamdsc/smart-sanskrit-annotator.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.874}
}

@InProceedings{oka-EtAl:2020:LREC,
  author    = {Oka, Teruaki  and  Ishimoto, Yuichi  and  Yagi,
Yutaka  and  Nakamura, Takenori  and  Asahara, Masayuki  and
Maekawa, Kikuo  and  Ogiso, Toshinobu  and  Koiso, Hanae  and
Sakoda, Kumiko  and  Kibe, Nobuko},
  title     = {KOTONOHA: A Corpus Concordance System for Skewer-
Searching NINJAL Corpora},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7077--7083},
  abstract  = {The National Institute for Japanese Language and
Linguistics, Japan (NINJAL, Japan), has developed several types of
corpora. For each corpus NINJAL provided an online search
environment, `Chunagon', which is a morphological-information-
annotation-based concordance system made publicly available in 2011.
NINJAL has now provided a skewer-search system `Kotonoha' based on
the `Chunagon' systems. This system enables querying of multiple
corpora by certain categories, such as register type and period.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.875}
}

@InProceedings{ogawa-EtAl:2020:LREC,
  author    = {Ogawa, Haruna  and  Nishikawa, Hitoshi  and
Tokunaga, Takenobu  and  Yokono, Hikaru},
  title     = {Gamification Platform for Collecting Task-oriented
Dialogue Data},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
```

```
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7084--7093},
  abstract  = {Demand for massive language resources is increasing
as the data-driven approach has established a leading position in
Natural Language Processing. However, creating dialogue corpora is
still a difficult task due to the complexity of the human dialogue
structure and the diversity of dialogue topics. Though crowdsourcing
is majorly used to assemble such data, it presents problems such as
less-motivated workers. We propose a platform for collecting task-
oriented situated dialogue data by using gamification. Combining a
video game with data collection benefits such as motivating workers
and cost reduction. Our platform enables data collectors to create
their original video game in which they can collect dialogue data of
various types of tasks by using the logging function of the
platform. Also, the platform provides the annotation function that
enables players to annotate their own utterances. The annotation can
be gamified aswell. We aim at high-quality annotation by introducing
such self-annotation method. We implemented a prototype of the
proposed platform and conducted a preliminary evaluation to obtain
promising results in terms of both dialogue data collection and
self-annotation.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.876}
}

@InProceedings{rose:2020:LREC,
  author    = {Rose, Ralph},
  title     = {Improving the Production Efficiency and Well-
formedness of Automatically-Generated Multiple-Choice Cloze
Vocabulary Questions},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7094--7101},
  abstract  = {Multiple-choice cloze (fill-in-the-blank) questions
are widely used in knowledge testing and are commonly used for
testing vocabulary knowledge. Word Quiz Constructor (WQC) is a Java
application that is designed to produce such test items
automatically from the Academic Word List (Coxhead, 2000) and using
various online and offline resources. The present work evaluates
recently added features of WQC to see whether they improve the
production quality and well-formedness of vocabulary quiz items over
previously implemented features in WQC. Results of a production test
and a well-formedness survey using Amazon Mechanical Turk show that
newly-introduced features (Linsear Write readability formula and
Google Books NGrams frequency list) significantly improve the
production quality of items over previous features (Automated
Readability Index and frequency list derived from the British
Academic Written English corpus). Items are produced faster and stem
```

sentences are shorter in length without any degradation in their well-formedness. Approximately 90\% of such items are judged well-formed, surpassing the rate of manually-produced items.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.877}
}

@InProceedings{rehbein-ruppenhofer-schmidt:2020:LREC,
  author   = {Rehbein, Ines and Ruppenhofer, Josef and Schmidt, Thomas},
  title   = {Improving Sentence Boundary Detection for Spoken Language Transcripts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7102--7111},
  abstract = {This paper presents experiments on sentence boundary detection in transcripts of spoken dialogues. Segmenting spoken language into sentence-like units is a challenging task, due to disfluencies, ungrammatical or fragmented structures and the lack of punctuation. In addition, one of the main bottlenecks for many NLP applications for spoken language is the small size of the training data, as the transcription and annotation of spoken language is by far more time-consuming and labour-intensive than processing written language. We therefore investigate the benefits of data expansion and transfer learning and test different ML architectures for this task. Our results show that data expansion is not straightforward and even data from the same domain does not always improve results. They also highlight the importance of modelling, i.e. of finding the best architecture and data representation for the task at hand. For the detection of boundaries in spoken language transcripts, we achieve a substantial improvement when framing the boundary detection problem assentence pair classification task, as compared to a sequence tagging approach.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.878}
}

@InProceedings{eskander-EtAl:2020:LREC,
  author   = {Eskander, Ramy and Callejas, Francesca and Nichols, Elizabeth and Klavans, Judith and Muresan, Smaranda},
  title   = {MorphAGram, Evaluation and Framework for Unsupervised Morphological Segmentation},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7112--7122},
  abstract = {Computational morphological segmentation has been an active research topic for decades as it is beneficial for many natural language processing tasks. With the high cost of manually

labeling data for morphology and the increasing interest in low-resource languages, unsupervised morphological segmentation has become essential for processing a typologically diverse set of languages, whether high-resource or low-resource. In this paper, we present and release MorphAGram, a publicly available framework for unsupervised morphological segmentation that uses Adaptor Grammars (AG) and is based on the work presented by Eskander et al. (2016). We conduct an extensive quantitative and qualitative evaluation of this framework on 12 languages and show that the framework achieves state-of-the-art results across languages of different typologies (from fusional to polysynthetic and from high-resource to low-resource).},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.879}
}

@InProceedings{rojowiec-roth-fink:2020:LREC,
  author   = {Rojowiec, Robin and Roth, Benjamin and Fink, Maximilian},
  title   = {Intent Recognition in Doctor-Patient Interviews},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {702--709},
  abstract = {Learning to interview patients to find out their disease is an essential part of the training of medical students. The practical part of this training has traditionally relied on paid actors that play the role of a patient to be interviewed. This process is expensive and severely limits the amount of practice per student. In this work, we present a novel data set and methods based on Natural Language Processing, for making progress towards modern applications and e-learning tools that support this training by providing language-based user interfaces with virtual patients. A data set of german transcriptions from live doctor-patient interviews was collected. These transcriptions are based on audio recordings of exercise sessions within the university and only the doctor's utterances could be transcribed. We annotated each utterance with an intent inventory characterizing the purpose of the question or statement. For some intent classes, the data only contains a few samples, and we apply Information Retrieval and Deep Learning methods that are robust with respect to small amounts of training data for recognizing the intent of an utterance and providing the correct response. Our results show that the models are effective and they provide baseline performance scores on the data set for further research.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.88}
}

@InProceedings{okinina-frey-weiss:2020:LREC,
  author   = {Okinina, Nadezda and Frey, Jennifer-Carmen and Weiss, Zarah},
  title   = {CTAP for Italian: Integrating Components for the

Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7123--7131},
  abstract = {Linguistic complexity research being a very actively developing field, an increasing number of text analysis tools are created that use natural language processing techniques for the automatic extraction of quantifiable measures of linguistic complexity. While most tools are designed to analyse only one language, the CTAP open source linguistic complexity measurement tool is capable of processing multiple languages, making cross-lingual comparisons possible. Although it was originally developed for English, the architecture has been ex-tended to support multi-lingual analyses. Here we present the Italian component of CTAP, describe its implementation and compare it to the existing linguistic complexity tools for Italian. Offering general text length statistics and features for lexical, syntactic, and morpho-syntactic complexity (including measures of lexical frequency, lexical diversity, lexical and syntactical variation, part-of-speech density), CTAP is currently the most comprehensive linguistic complexity measurement tool for Italian and the only one allowing the comparison of Italian texts to multiple other languages within one tool.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.880}
}

@InProceedings{stodden-qasemizadeh-kallmeyer:2020:LREC,
  author   = {Stodden, Regina and QasemiZadeh, Behrang and Kallmeyer, Laura},
  title   = {Do you Feel Certain about your Annotation? A Web-based Semantic Frame Annotation Tool Considering Annotators' Concerns and Behaviors},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7132--7139},
  abstract = {In this system demonstration paper, we present an open-source web-based application with a responsive design for modular semantic frame annotation (SFA). Besides letting experienced and inexperienced users do suggestion-based and slightly-controlled annotations, the system keeps track of the time and changes during the annotation process and stores the users' confidence with the current annotation. This collected metadata can be used to get insights regarding the difficulty of an annotation with the same type or frame or can be used as an input of an annotation cost measurement for an active learning algorithm. The tool was already

used to build a manually annotated corpus with semantic frames and its arguments for task 2 of SemEval 2019 regarding unsupervised lexical frame induction (QasemiZadeh et al., 2019). Although English sentences from the Wall Street Journal corpus of the Penn Treebank were annotated for this task, it is also possible to use the proposed tool for the annotation of sentences in other languages.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.881}
}

@InProceedings{qader-portet-labbe:2020:LREC,
  author   = {Qader, Raheel and Portet, François and Labbe, Cyril},
  title   = {Seq2SeqPy: A Lightweight and Customizable Toolkit for Neural Sequence-to-Sequence Modeling},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7140--7144},
  abstract = {We present Seq2SeqPy a lightweight toolkit for sequence-to-sequence modeling that prioritizes simplicity and ability to customize the standard architectures easily. The toolkit supports several known architectures such as Recurrent Neural Networks, Pointer Generator Networks, and transformer model. We evaluate the toolkit on two datasets and we show that the toolkit performs similarly or even better than a very widely used sequence-to-sequence toolkit.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.882}
}

@InProceedings{brunato-EtAl:2020:LREC,
  author   = {Brunato, Dominique and Cimino, Andrea and Dell'Orletta, Felice and Venturi, Giulia and Montemagni, Simonetta},
  title   = {Profiling-UD: a Tool for Linguistic Profiling of Texts},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7145--7151},
  abstract = {In this paper, we introduce Profiling-UD, a new text analysis tool inspired to the principles of linguistic profiling that can support language variation research from different perspectives. It allows the extraction of more than 130 features, spanning across different levels of linguistic description. Beyond the large number of features that can be monitored, a main novelty of Profiling-UD is that it has been specifically devised to be multilingual since it is based on the Universal Dependencies framework. In the second part of the paper, we demonstrate the

effectiveness of these features in a number of theoretical and applicative studies in which they were successfully used for text and author profiling.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.883}
}

@InProceedings{laur-EtAl:2020:LREC,
  author   = {Laur, Sven and Orasmaa, Siim and Särg, Dage and Tammo, Paul},
  title   = {EstNLTK 1.6: Remastered Estonian NLP Pipeline},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7152--7160},
  abstract = {The goal of the EstNLTK Python library is to provide a unified programming interface for natural language processing in Estonian. As such, previous versions of the library have been immensely successful both in academic and industrial circles. However, they also contained serious structural limitations -- it was hard to add new components and there was a lack of fine-grained control needed for back-end programming. These issues have been explicitly addressed in the EstNLTK library while preserving the intuitive interface for novices. We have remastered the basic NLP pipeline by adding many data cleaning steps that are necessary for analyzing real-life texts, and state of the art components for morphological analysis and fact extraction. Our evaluation on unlabelled data shows that the remastered basic NLP pipeline outperforms both the previous version of the toolkit, as well as neural models of StanfordNLP. In addition, EstNLTK contains a new interface for storing, processing and querying text objects in Postgres database which greatly simplifies processing of large text collections. EstNLTK is freely available under the GNU GPL version 2 license, which is standard for academic software.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.884}
}

@InProceedings{chiarcos-glaser:2020:LREC,
  author   = {Chiarcos, Christian and Glaser, Luis},
  title   = {A Tree Extension for CoNLL-RDF},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7161--7169},
  abstract = {The technological bridges between knowledge graphs and natural language processing are of utmost importance for the future development of language technology. CoNLL-RDF is a technology that provides such a bridge for popular one-word-per-line formats as widely used in NLP (e.g., the CoNLL Shared Tasks), annotation

(Universal Dependencies, Unimorph), corpus linguistics (Corpus WorkBench, CWB) and digital lexicography (SketchEngine): Every empty-line separated table (usually a sentence) is parsed into an graph, can be freely manipulated and enriched using W3C-standardized RDF technology, and then be serialized back into in a TSV format, RDF or other formats. An important limitation is that CoNLL-RDF provides native support for word-level annotations only. This does include dependency syntax and semantic role annotations, but neither phrase structures nor text structure. We describe the extension of the CoNLL-RDF technology stack for two vocabulary extensions of CoNLL-TSV, the PTB bracket notation used in earlier CoNLL Shared Tasks and the extension with XML markup elements featured by CWB and SketchEngine. In order to represent the necessary extensions of the CoNLL vocabulary in an adequate fashion, we employ the POWLA vocabulary for representing and navigating in tree structures.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.885}
}

@InProceedings{trips-percillier:2020:LREC,
  author   = {Trips, Carola and Percillier, Michael},
  title   = {Lemmatising Verbs in Middle English Corpora: The Benefit of Enriching the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), the Parsed Corpus of Middle English Poetry (PCMEP), and A Parsed Linguistic Atlas of Early Middle English (PLAEME)},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7170--7178},
  abstract = {This paper describes the lemmatisation of three annotated corpora of Middle English—the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), the Parsed Corpus of Middle English Poetry (PCMEP), and A Parsed Linguistic Atlas of Early Middle English (PLAEME) – which is a prerequisite for systematically investigating the argument structures of verbs of the given time. Creating this tool and enriching existing parsed corpora of Middle English is part of the project Borrowing of Argument Structure in Contact Situations (BASICS) which seeks to explain to which extent verbs copied from Old French had an impact on the grammar of Middle English. First, we lemmatised the PPCME2 by (1) creating an inventory of form-lemma correspondences linking forms in the PPCME2 to lemmas in the MED, and (2) inserting this lemma information into the corpus (precision: 94.85\%, recall: 98.92\%). Second, we enriched the PCMEP and PLAEME, which adopted the annotation format of the PPCME2, with verb lemmas to undertake studies that fill the well-known data gap in the subperiod (1250–1350) of the PPCME2. The case study of reflexives shows that with our method we gain much more reliable results in terms of diachrony, diatopy and contact-induced change.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.886}
}

```
@InProceedings{stajner-nisioi-hulpu:2020:LREC,
  author    = {Stajner, Sanja  and  Nisioi, Sergiu  and  Hulpuș,
Ioana},
  title     = {CoCo: A Tool for Automatically Assessing Conceptual
Complexity of Texts},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7179--7186},
  abstract  = {Traditional text complexity assessment usually takes
into account only syntactic and lexical text complexity. The task of
automatic assessment of conceptual text complexity, important for
maintaining reader’s interest and text adaptation for struggling
readers, has only been proposed recently. In this paper, we present
CoCo - a tool for automatic assessment of conceptual text
complexity, based on using the current state-of-the-art unsupervised
approach. We make the code and API freely available for research
purposes, and describe the code and the possibility for its
personalization and adaptation in details. We compare the current
implementation with the state of the art, discussing the influence
of the choice of entity linker on the performances of the tool.
Finally, we present results obtained on two widely used text
simplification corpora, discussing the full potential of the tool.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.887}
}

@InProceedings{verner-vernerov:2020:LREC,
  author    = {Verner, Jonathan  and  Vernerová, Anna},
  title     = {PyVallex: A Processing System for Valency Lexicon
Data},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7187--7193},
  abstract  = {PyVallex is a Python-based system for presenting,
searching/filtering, editing/extending and automatic processing of
machine-readable lexicon data originally available in a text-based
format. The system consists of several components: a parser for the
specific lexicon format used in several valency lexicons, a data-
validation framework, a regular expression based search engine, a
map-reduce style framework for querying the lexicon data and a web-
based interface integrating complex search and some basic editing
capabilities. PyVallex provides most of the typical functionalities
of a Dictionary Writing System (DWS), such as multiple presentation
modes for the underlying lexical database, automatic evaluation of
consistency tests, and a mechanism of merging updates coming from
multiple sources. The editing functionality is currently limited to
```

```
the client-side interface and edits of existing lexical entries, but
additional script-based operations on the database are also
possible. The code is published under the open source MIT license
and is also available in the form of a Python module for integrating
into other software.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.888}
}

@InProceedings{fiorelli-EtAl:2020:LREC,
  author    = {Fiorelli, Manuel  and  Stellato, Armando  and
Lorenzetti, Tiziano  and  Turbati, Andrea  and  Schmitz, Peter  and
Francesconi, Enrico  and  Hajlaoui, Najeh  and  Batouche, Brahim},
  title     = {Editing OntoLex-Lemon in VocBench 3},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7194--7203},
  abstract  = {OntoLex-Lemon is a collection of RDF vocabularies for
specifying the verbalization of ontologies in natural language.
Beyond its original scope, OntoLex-Lemon, as well as its predecessor
Monnet lemon, found application in the Linguistic Linked Open Data
cloud to represent and interlink language resources on the Semantic
Web. Unfortunately, generic ontology and RDF editors were considered
inconvenient to use with OntoLex-Lemon because of its complex design
patterns and other peculiarities, including indirection, reification
and subtle integrity constraints. This perception led to the
development of dedicated editors, trading the flexibility of RDF in
combining different models (and the features already available in
existing RDF editors) for a more direct and streamlined editing of
OntoLex-Lemon patterns. In this paper, we investigate on the
benefits gained by extending an already existing RDF editor,
VocBench 3, with capabilities closely tailored to OntoLex-Lemon and
on the challenges that such extension implies. The outcome of such
investigation is twofold: a vertical assessment of a new editor for
OntoLex-Lemon and, in the broader scope of RDF editor design, a new
perspective on which flexibility and extensibility characteristics
an editor should meet in order to cover new core modeling
vocabularies, for which OntoLex-Lemon represents a use case.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.889}
}

@InProceedings{hmamouche-EtAl:2020:LREC,
  author    = {Hmamouche, Youssef  and  Prévot, Laurent  and  Ochs,
Magalie  and  Chaminade, Thierry},
  title     = {BrainPredict: a Tool for Predicting and Visualising
Local Brain Activity},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
```

```
  publisher      = {European Language Resources Association},
  pages     = {710--716},
  abstract  = {In this paper, we present a tool allowing dynamic
prediction and visualization of an individual's local brain activity
during a conversation. The prediction module of this tool is based
on classifiers trained using a corpus of human-human and human-robot
conversations including fMRI recordings. More precisely, the module
takes as input behavioral features computed from raw data, mainly
the participant and the interlocutor speech but also the
participant's visual input and eye movements. The visualisation
module shows in real-time the dynamics of brain active areas
synchronised with the behavioral raw data. In addition, it shows
which integrated behavioral features are used to predict the
activity in individual brain areas.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.89}
}

@InProceedings{forti-EtAl:2020:LREC,
  author    = {Forti, Luciana  and  Grego Bolli, Giuliana  and
Santarelli, Filippo  and  Santucci, Valentino  and  Spina,
Stefania},
  title     = {MALT-IT2: A New Resource to Measure Text Difficulty
in Light of CEFR Levels for Italian L2 Learning},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7204--7211},
  abstract  = {This paper presents a new resource for automatically
assessing text difficulty in the context of Italian as a second or
foreign language learning and teaching. It is called MALT-IT2, and
it automatically classifies inputted texts according to the CEFR
level they are more likely to belong to. After an introduction to
the field of automatic text difficulty assessment, and an overview
of previous related work, we describe the rationale of the project,
the corpus and computational system it is based on. Experiments were
conducted in order to investigate the reliability of the system. The
results show that the system is able to obtain a good prediction
accuracy, while a further analysis was conducted in order to
identify the categories of features which mostly influenced the
predictions.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.890}
}

@InProceedings{fth-EtAl:2020:LREC,
  author    = {Fäth, Christian  and  Chiarcos, Christian  and
Ebbrecht, Björn  and  Ionov, Maxim},
  title     = {Fintan - Flexible, Integrated Transformation and
Annotation eNgineering},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
```

```
  year          = {2020},
  address       = {Marseille, France},
  publisher     = {European Language Resources Association},
  pages     = {7212--7221},
  abstract  = {We introduce the Flexible and Integrated
Transformation and Annotation eNgeneering (Fintan) platform for
converting heterogeneous linguistic resources to RDF. With its
modular architecture, workflow management and visualization
features, Fintan facilitates the development of complex
transformation pipelines by integrating generic RDF converters and
augmenting them with extended graph processing capabilities:
Existing converters can be easily deployed to the system by means of
an ontological data structure which renders their properties and the
dependencies between transformation steps. Development of subsequent
graph transformation steps for resource transformation, annotation
engineering or entity linking is further facilitated by a novel
visual rendering of SPARQL queries. A graphical workflow manager
allows to easily manage the converter modules and combine them to
new transformation pipelines. Employing the stream-based graph
processing approach first implemented with CoNLL-RDF, we address
common challenges and scalability issues when transforming resources
and showcase the performance of Fintan by means of a purely graph-
based transformation of the Universal Morphology data to RDF.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.891}
}

@InProceedings{waszczuk-EtAl:2020:LREC,
  author    = {Waszczuk, Jakub  and  Wang, Ilaine  and  Antoine,
Jean-Yves  and  Halftermeyer, Anaïs},
  title     = {Contemplata, a Free Platform for Constituency
Treebank Annotation},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {7222--7229},
  abstract  = {This paper describes Contemplata, an annotation
platform that offers a generic solution for treebank building as
well as treebank enrichment with relations between syntactic nodes.
Contemplata is dedicated to the annotation of constituency trees.
The framework includes support for syntactic parsers, which provide
automatic annotations to be manually revised. The balanced strategy
of annotation between automatic parsing and manual revision allows
to reduce the annotator workload, which favours data reliability.
The paper presents the software architecture of Contemplata,
describes its practical use and eventually gives two examples of
annotation projects that were conducted on the platform.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.892}
}

@InProceedings{rim-EtAl:2020:LREC2,
  author    = {Rim, Kyeongmin  and  Lynch, Kelley  and  Verhagen,
```

Marc and Ide, Nancy and Pustejovsky, James},
  title   = {Interchange Formats for Visualization: LIF and MMIF},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7230--7237},
  abstract = {Promoting interoperrable computational linguistics (CL) and natural language processing (NLP) application platforms and interchange-able data formats have contributed improving discoverabilty and accessbility of the openly available NLP software. In this paper, wediscuss the enhanced data visualization capabilities that are also enabled by inter-operating NLP pipelines and interchange formats.For adding openly available visualization tools and graphical annotation tools to the Language Applications Grid (LAPPS Grid) andComputational Linguistics Applications for Multimedia Services (CLAMS) toolboxes, we have developed interchange formats that cancarry annotations and metadata for text and audiovisual source data. We descibe those data formats and present case studies where wesuccessfully adopt open-source visualization tools and combine them with CL tools.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.893}
}

@InProceedings{davidson-EtAl:2020:LREC,
  author   = {Davidson, Sam and Yamada, Aaron and Fernandez Mira, Paloma and Carando, Agustina and Sanchez Gutierrez, Claudia H. and Sagae, Kenji},
  title   = {Developing NLP Tools with a New Corpus of Learner Spanish},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7238--7243},
  abstract = {The development of effective NLP tools for the L2 classroom depends largely on the availability of large annotated corpora of language learner text. While annotated learner corpora of English are widely available, large learner corpora of Spanish are less common. Those Spanish corpora that are available do not contain the annotations needed to facilitate the development of tools beneficial to language learners, such as grammatical error correction. As a result, the field has seen little research in NLP tools designed to benefit Spanish language learners and teachers. We introduce COWS-L2H, a freely available corpus of Spanish learner data which includes error annotations and parallel corrected text to help researchers better understand L2 development, to examine teaching practices empirically, and to develop NLP tools to better serve the Spanish teaching community. We demonstrate the utility of this corpus by developing a neural-network based grammatical error

correction system for Spanish learner writing.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.894}
}

@InProceedings{rodrigues-EtAl:2020:LREC,
  author   = {Rodrigues, Francisco and Lima, Rinaldo and Domingues, William and Fidalgo, Robson and Chifu, Adrian and Espinasse, Bernard and Fournier, Sébastien},
  title   = {DeepNLPF: A Framework for Integrating Third Party NLP Tools},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7244--7251},
  abstract = {Natural Language Processing (NLP) of textual data is usually broken down into a sequence of several subtasks, where the output of one the subtasks becomes the input to the following one, which constitutes an NLP pipeline. Many third-party NLP tools are currently available, each performing distinct NLP subtasks. However, it is difficult to integrate several NLP toolkits into a pipeline due to many problems, including different input/output representations or formats, distinct programming languages, and tokenization issues. This paper presents DeepNLPF, a framework that enables easy integration of third-party NLP tools, allowing the user to preprocess natural language texts at lexical, syntactic, and semantic levels. The proposed framework also provides an API for complete pipeline customization including the definition of input/output formats, integration plugin management, transparent ultiprocessing execution strategies, corpus-level statistics, and database persistence. Furthermore, the DeepNLPF user-friendly GUI allows its use even by a non-expert NLP user. We conducted runtime performance analysis showing that DeepNLPF not only easily integrates existent NLP toolkits but also reduces significant runtime processing compared to executing the same NLP pipeline in a sequential manner.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.895}
}

@InProceedings{aralikatte-sgaard:2020:LREC,
  author   = {Aralikatte, Rahul and Søgaard, Anders},
  title   = {Model-based Annotation of Coreference},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {74--79},
  abstract = {Humans do not make inferences over texts, but over models of what texts are about. When annotators are asked to annotate coreferent spans of text, it is therefore a somewhat

unnatural task. This paper presents an alternative in which we preprocess documents, linking entities to a knowledge base, and turn the coreference annotation task -- in our case limited to pronouns -- into an annotation task where annotators are asked to assign pronouns to entities. Model-based annotation is shown to lead to faster annotation and higher inter-annotator agreement, and we argue that it also opens up an alternative approach to coreference resolution. We present two new coreference benchmark datasets, for English Wikipedia and English teacher-student dialogues, and evaluate state-of-the-art coreference resolvers on them.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.9}
}

@InProceedings{senese-EtAl:2020:LREC,
  author   = {Senese, Matteo Antonio and Rizzo, Giuseppe and Dragoni, Mauro and Morisio, Maurizio},
  title   = {MTSI-BERT: A Session-aware Knowledge-based Conversational Agent},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {717--725},
  abstract = {In the last years, the state of the art of NLP research has made a huge step forward. Since the release of ELMo (Peters et al., 2018), a new race for the leading scoreboards of all the main linguistic tasks has begun. Several models have been published achieving promising results in all the major NLP applications, from question answering to text classification, passing through named entity recognition. These great research discoveries coincide with an increasing trend for voice-based technologies in the customer care market. One of the next biggest challenges in this scenario will be the handling of multi-turn conversations, a type of conversations that differs from single-turn by the presence of multiple related interactions. The proposed work is an attempt to exploit one of these new milestones to handle multi-turn conversations. MTSI-BERT is a BERT-based model achieving promising results in intent classification, knowledge base action prediction and end of dialogue session detection, to determine the right moment to fulfill the user request. The study about the realization of PuffBot, an intelligent chatbot to support and monitor people suffering from asthma, shows how this type of technique could be an important piece in the development of future chatbots.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.90}
}

@InProceedings{georgila-EtAl:2020:LREC1,
  author   = {Georgila, Kallirroi and Gordon, Carla and Yanov, Volodymyr and Traum, David},
  title   = {Predicting Ratings of Real Dialogue Participants from Artificial Data and Ratings of Human Dialogue Observers},

```
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {726--734},
  abstract  = {We collected a corpus of dialogues in a Wizard of Oz
(WOz) setting in the Internet of Things (IoT) domain. We asked users
participating in these dialogues to rate the system on a number of
aspects, namely, intelligence, naturalness, personality,
friendliness, their enjoyment, overall quality, and whether they
would recommend the system to others. Then we asked dialogue
observers, i.e., Amazon Mechanical Turkers (MTurkers), to rate these
dialogues on the same aspects. We also generated simulated dialogues
between dialogue policies and simulated users and asked MTurkers to
rate them again on the same aspects. Using linear regression, we
developed dialogue evaluation functions based on features from the
simulated dialogues and the MTurkers' ratings, the WOz dialogues and
the MTurkers' ratings, and the WOz dialogues and the WOz
participants' ratings. We applied all these dialogue evaluation
functions to a held-out portion of our WOz dialogues, and we report
results on the predictive power of these different types of dialogue
evaluation functions. Our results suggest that for three
conversational aspects (intelligence, naturalness, overall quality)
just training evaluation functions on simulated data could be
sufficient.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.91}
}

@InProceedings{alavi-leuski-traum:2020:LREC,
  author    = {Alavi, Seyed Hossein  and  Leuski, Anton  and  Traum,
David},
  title     = {Which Model Should We Use for a Real-World
Conversational Dialogue System? a Cross-Language Relevance Model or
a Deep Neural Net?},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {735--742},
  abstract  = {We compare two models for corpus-based selection of
dialogue responses: one based on cross-language relevance with a
cross-language LSTM model. Each model is tested on multiple corpora,
collected from two different types of dialogue source material.
Results show that while the LSTM model performs adequately on a very
large corpus (millions of utterances), its performance is dominated
by the cross-language relevance model for a more moderate-sized
corpus (ten thousands of utterances).},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.92}
}
```

```
@InProceedings{kontogiorgos-sibirtseva-gustafson:2020:LREC,
  author    = {Kontogiorgos, Dimosthenis  and  Sibirtseva, Elena
and  Gustafson, Joakim},
  title     = {Chinese Whispers: A Multimodal Dataset for Embodied
Language Grounding},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {743--749},
  abstract  = {In this paper, we introduce a multimodal dataset in
which subjects are instructing each other how to assemble IKEA
furniture. Using the concept of `Chinese Whispers', an old
children's game, we employ a novel method to avoid implicit
experimenter biases. We let subjects instruct each other on the
nature of the task: the process of the furniture assembly.
Uncertainty, hesitations, repairs and self-corrections are naturally
introduced in the incremental process of establishing common ground.
The corpus consists of 34 interactions, where each subject first
assembles and then instructs. We collected speech, eye-gaze,
pointing gestures, and object movements, as well as subjective
interpretations of mutual understanding, collaboration and task
recall. The corpus is of particular interest to researchers who are
interested in multimodal signals in situated dialogue, especially in
referential communication and the process of language grounding.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.93}
}

@InProceedings{kumar-EtAl:2020:LREC2,
  author    = {Kumar, Gaurav  and  Joshi, Rishabh  and  Singh,
Jaspreet  and  Yenigalla, Promod},
  title     = {AMUSED: A Multi-Stream Vector Representation Method
for Use in Natural Dialogue},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {750--758},
  abstract  = {The problem of building a coherent and non-monotonous
conversational agent with proper discourse and coverage is still an
area of open research. Current architectures only take care of
semantic and contextual information for a given query and fail to
completely account for syntactic and external knowledge which are
crucial for generating responses in a chit-chat system. To overcome
this problem, we propose an end to end multi-stream deep learning
architecture that learns unified embeddings for query-response pairs
by leveraging contextual information from memory networks and
syntactic information by incorporating Graph Convolution Networks
(GCN) over their dependency parse. A stream of this network also
utilizes transfer learning by pre-training a bidirectional
```

transformer to extract semantic representation for each input sentence and incorporates external knowledge through the neighborhood of the entities from a Knowledge Base (KB). We benchmark these embeddings on the next sentence prediction task and significantly improve upon the existing techniques. Furthermore, we use AMUSED to represent query and responses along with its context to develop a retrieval based conversational agent which has been validated by expert linguists to have comprehensive engagement with humans.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.94}
}

@InProceedings{somashekarappa-howes-sayeed:2020:LREC,
  author   = {Somashekarappa, Vidya and Howes, Christine and Sayeed, Asad},
  title   = {An Annotation Approach for Social and Referential Gaze in Dialogue},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {759--765},
  abstract = {This paper introduces an approach for annotating eye gaze considering both its social and the referential functions in multi-modal human-human dialogue. Detecting and interpreting the temporal patterns of gaze behavior cues is natural for humans and also mostly an unconscious process. However, these cues are difficult for conversational agents such as robots or avatars to process or generate. The key factor is to recognize these variants and carry out a successful conversation, as misinterpretation can lead to total failure of the given interaction. This paper introduces an annotation scheme for eye-gaze in human-human dyadic interactions that is intended to facilitate the learning of eye-gaze patterns in multi-modal natural dialogue.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.95}
}

@InProceedings{booth-EtAl:2020:LREC1,
  author   = {Booth, Hannah and Breitbarth, Anne and Ecay, Aaron and Farasyn, Melissa},
  title   = {A Penn-style Treebank of Middle Low German},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {766--775},
  abstract = {We outline the issues and decisions involved in creating a Penn-style treebank of Middle Low German (MLG, 1200-1650), which will form part of the Corpus of Historical Low German (CHLG). The attestation for MLG is rich, but the syntax of

the language remains relatively understudied. The development of a syntactically annotated corpus for the language will facilitate future studies with a strong empirical basis, building on recent work which indicates that, syntactically, MLG occupies a position in its own right within West Germanic. In this paper, we describe the background for the corpus and the process by which texts were selected to be included. In particular, we focus on the decisions involved in the syntactic annotation of the corpus, specifically, the practical and linguistic reasons for adopting the Penn annotation scheme, the stages of the annotation process itself, and how we have adapted the Penn scheme for syntactic features specific to MLG. We also discuss the issue of data uncertainty, which is a major issue when building a corpus of an under-researched language stage like MLG, and some novel ways in which we capture this uncertainty in the annotation.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.96}
}

@InProceedings{hazem-EtAl:2020:LREC,
  author   = {Hazem, Amir and Daille, Beatrice and Kermorvant, Christopher and Stutzmann, Dominique and Bonhomme, Marie-Laurence and Maarand, Martin and Boillet, Mélodie},
  title   = {Books of Hours. the First Liturgical Data Set for Text Segmentation.},
  booktitle   = {Proceedings of The 12th Language Resources and Evaluation Conference},
  month   = {May},
  year   = {2020},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {776--784},
  abstract = {The Book of Hours was the bestseller of the late Middle Ages and Renaissance. It is a historical invaluable treasure, documenting the devotional practices of Christians in the late Middle Ages. Up to now, its textual content has been scarcely studied because of its manuscript nature, its length and its complex content. At first glance, it looks too standardized. However, the study of book of hours raises important challenges: (i) in image analysis, its often lavish ornamentation (illegible painted initials, line-fillers, etc.), abbreviated words, multilingualism are difficult to address in Handwritten Text Recognition (HTR); (ii) its hierarchical entangled structure offers a new field of investigation for text segmentation; (iii) in digital humanities, its textual content gives opportunities for historical analysis. In this paper, we provide the first corpus of books of hours, which consists of Latin transcriptions of 300 books of hours generated by Handwritten Text Recognition (HTR) - that is like Optical Character Recognition (OCR) but for handwritten and not printed texts. We designed a structural scheme of the book of hours and annotated manually two books of hours according to this scheme. Lastly, we performed a systematic evaluation of the main state of the art text segmentation approaches.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.97}
}

```
@InProceedings{zinin-xu:2020:LREC,
  author    = {Zinin, Sergey  and  Xu, Yang},
  title     = {Corpus of Chinese Dynastic Histories: Gender Analysis
over Two Millennia},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {785--793},
  abstract  = {Chinese dynastic histories form a large continuous
linguistic space of approximately 2000 years, from the 3rd century
BCE to the 18th century CE. The histories are documented in
Classical (Literary) Chinese in a corpus of over 20 million
characters, suitable for the computational analysis of historical
lexicon and semantic change. However, there is no freely available
open-source corpus of these histories, making Classical Chinese low-
resource. This project introduces a new open-source corpus of
twenty-four dynastic histories covered by Creative Commons license.
An original list of Classical Chinese gender-specific terms was
developed as a case study for analyzing the historical linguistic
use of male and female terms. The study demonstrates considerable
stability in the usage of these terms, with dominance of male terms.
Exploration of word meanings uses keyword analysis of focus corpora
created for gender-specific terms. This method yields meaningful
semantic representations that can be used for future studies of
diachronic semantics.},
  url       = {https://www.aclweb.org/anthology/2020.lrec-1.98}
}

@InProceedings{fischer-EtAl:2020:LREC,
  author    = {Fischer, Stefan  and  Knappen, Jörg  and  Menzel,
Katrin  and  Teich, Elke},
  title     = {The Royal Society Corpus 6.0: Providing 300+ Years of
Scientific Writing for Humanistic Study},
  booktitle      = {Proceedings of The 12th Language Resources and
Evaluation Conference},
  month          = {May},
  year           = {2020},
  address        = {Marseille, France},
  publisher      = {European Language Resources Association},
  pages     = {794--802},
  abstract  = {We present a new, extended version of the Royal
Society Corpus (RSC), a diachronic corpus of scientific English now
covering 300+ years of scientific writing (1665–1996). The corpus
comprises 47 837 texts, primarily scientific articles, and is based
on publications of the Royal Society of London, mainly its
Philosophical Transactions and Proceedings. The corpus has been
built on the basis of the FAIR principles and is freely available
under a Creative Commons license, excluding copy-righted parts. We
provide information on how the corpus can be found, the file formats
available for download as well as accessibility via a web-based
```

corpus query platform. We show a number of analytic tools that we have implemented for better usability and provide an example of use of the corpus for linguistic analysis as well as examples of subsequent, external uses of earlier releases. We place the RSC against the background of existing English diachronic/scientific corpora, elaborating on its value for linguistic and humanistic study.},
  url   = {https://www.aclweb.org/anthology/2020.lrec-1.99}
}