Language Resources and Evaluation Conference LREC 2022 Proceedings

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:2022,
  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 Jan Odijk and Stelios Piperidis},
  title   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  url   = {https://aclanthology.org/2022.lrec-1}
}

@InProceedings{costa-EtAl:2022:LREC,
  author   = {Costa, Alexandre Diniz da and Coutinho Marim, Mateus and Matos, Ely and Timponi Torrent, Tiago},
  title   = {Domain Adaptation in Neural Machine Translation using a Qualia-Enriched FrameNet},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1--12},
  abstract = {In this paper we present Scylla, a methodology for domain adaptation of Neural Machine Translation (NMT) systems that make use of a multilingual FrameNet enriched with qualia relations as an external knowledge base. Domain adaptation techniques used in NMT usually require fine-tuning and in-domain training data, which may pose difficulties for those working with lesser-resourced languages and may also lead to performance decay of the NMT system for out-of-domain sentences. Scylla does not require fine-tuning of the NMT model, avoiding the risk of model over-fitting and consequent decrease in performance for out-of-domain translations. Two versions of Scylla are presented: one using the source sentence as input, and another one using the target sentence. We evaluate Scylla in comparison to a state-of-the-art commercial NMT system in an experiment in which 50 sentences from the Sports domain are translated from Brazilian Portuguese to English. The two versions of Scylla significantly outperform the baseline commercial system in HTER.},
  url   = {https://aclanthology.org/2022.lrec-1.1}
}

@InProceedings{blc-can:2022:LREC,
  author   = {Bölücü, Necva and Can, Burcu},
  title   = {Turkish Universal Conceptual Cognitive Annotation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {89--99},
  abstract = {Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013a) is a cross-lingual semantic annotation framework that provides an easy annotation without any requirement for linguistic background. UCCA-annotated datasets have been already released in English, French, and German. In this paper, we introduce the first UCCA-annotated Turkish dataset that currently involves 50 sentences obtained from the METU-Sabanci Turkish Treebank (Atalay et al., 2003; Oflazeret al., 2003). We followed a semi-automatic annotation approach, where an external semantic parser is utilised for an initial annotation of the dataset, which is partially accurate and requires refinement. We manually revised the annotations obtained from the semantic parser that are not in line with the UCCA rules that we defined for Turkish. We used the same external semantic parser for evaluation purposes and conducted experiments with both zero-shot and few-shot learning. While the parser cannot predict remote edges in zero-shot setting, using even a small subset of training data in few-shot setting increased the overall F-1 score including the remote edges. This is the initial version of the annotated dataset and we are currently extending the dataset. We will release the current Turkish UCCA annotation guideline along with the annotated dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.10}
}

@InProceedings{gogoulou-EtAl:2022:LREC,
  author   = {Gogoulou, Evangelia and Ekgren, Ariel and Isbister, Tim and Sahlgren, Magnus},
  title   = {Cross-lingual Transfer of Monolingual Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {948--955},
  abstract = {Recent studies in cross-lingual learning using multilingual models have cast doubt on the previous hypothesis that shared vocabulary and joint pre-training are the keys to cross-lingual generalization. We introduce a method for transferring monolingual models to other languages through continuous pre-training and study the effects of such transfer from four different languages to English. Our experimental results on GLUE show that the transferred models outperform an English model trained from scratch, independently of the source language. After probing the model representations, we find that model knowledge from the source language enhances the learning of syntactic and semantic knowledge in English.},
  url   = {https://aclanthology.org/2022.lrec-1.100}
}

@InProceedings{petersenfrey-EtAl:2022:LREC,

  author   = {Petersen-Frey, Fynn and Soll, Marcus and Kobras, Louis and Johannsen, Melf and Kling, Peter and Biemann, Chris},
  title   = {Dataset of Student Solutions to Algorithm and Data Structure Programming Assignments},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {956--962},
  abstract = {We present a dataset containing source code solutions to algorithmic programming exercises solved by hundreds of Bachelor-level students at the University of Hamburg. These solutions were collected during the winter semesters 2019/2020, 2020/2021 and 2021/2022. The dataset contains a set of solutions to a total of 21 tasks written in Java as well as Python and a total of over 1500 individual solutions. All solutions were submitted through Moodle and the Coderunner plugin and passed a number of test cases (including randomized tests), such that they can be considered as working correctly. All students whose solutions are included in the dataset gave their consent into publishing their solutions. The solutions are pseudonymized with a random solution ID. Included in this paper is a short analysis of the dataset containing statistical data and highlighting a few anomalies (e.g. the number of solutions per task decreases for the last few tasks due to grading rules). We plan to extend the dataset with tasks and solutions from upcoming courses.},
  url   = {https://aclanthology.org/2022.lrec-1.101}
}

@InProceedings{mondal-EtAl:2022:LREC,
  author   = {Mondal, Ishani and Bali, Kalika and Jain, Mohit and Choudhury, Monojit and O'Neill, Jacki and Ochieng, Millicent and Awori, Kagnoya and Ronen, Keshet},
  title   = {Language Patterns and Behaviour of the Peer Supporters in Multilingual Healthcare Conversational Forums},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {963--975},
  abstract = {In this work, we conduct a quantitative linguistic analysis of the language usage patterns of multilingual peer supporters in two health-focused WhatsApp groups in Kenya comprising of youth living with HIV. Even though the language of communication for the group was predominantly English, we observe frequent use of Kiswahili, Sheng and code-mixing among the three languages. We present an analysis of language choice and its accommodation, different functions of code-mixing, and relationship between sentiment and code-mixing. To explore the effectiveness of off-the-

shelf Language Technologies (LT) in such situations, we attempt to build a sentiment analyzer for this dataset. Our experiments demonstrate the challenges of developing LT and therefore effective interventions for such forums and languages. We provide recommendations for language resources that should be built to address these challenges.},
  url   = {https://aclanthology.org/2022.lrec-1.102}
}

@InProceedings{yong-EtAl:2022:LREC,
  author   = {Yong, Zheng Xin and Watson, Patrick D. and Timponi Torrent, Tiago and Czulo, Oliver and Baker, Collin},
  title   = {Frame Shift Prediction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {976--986},
  abstract = {Frame shift is a cross-linguistic phenomenon in translation which results in corresponding pairs of linguistic material evoking different frames. The ability to predict frame shifts would enable (semi-)automatic creation of multilingual frame annotations and thus speeding up FrameNet creation through annotation projection. Here, we first characterize how frame shifts result from other linguistic divergences such as translational divergences and construal differences. Our analysis also shows that many pairs of frames in frame shifts are multi-hop away from each other in Berkeley FrameNet's net-like configuration. Then, we propose the Frame Shift Prediction task and demonstrate that our graph attention networks, combined with auxiliary training, can learn cross-linguistic frame-to-frame correspondence and predict frame shifts.},
  url   = {https://aclanthology.org/2022.lrec-1.103}
}

@InProceedings{bigi-zimmermann-andr:2022:LREC,
  author   = {BIGI, Brigitte and Zimmermann, Maryvonne and André, Carine},
  title   = {CLeLfPC: a Large Open Multi-Speaker Corpus of French Cued Speech},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {987--994},
  abstract = {Cued Speech is a communication system developed for deaf people to complement speechreading at the phonetic level with hands. This visual communication mode uses handshapes in different placements near the face in combination with the mouth movements of speech to make the phonemes of spoken language look different from

each other. This paper describes CLeLfPC - Corpus de Lecture en Langue française Parlée Complétée, a corpus of French Cued Speech. It consists in about 4 hours of audio and HD video recordings of 23 participants. The recordings are 160 different isolated 'CV' syllables repeated 5 times, 320 words or phrases repeated 2-3 times and about 350 sentences repeated 2-3 times. The corpus is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It can be used for any further research or teaching purpose. The corpus includes orthographic transliteration and other phonetic annotations on 5 of the recorded topics, i.e. syllables, words, isolated sentences and a text. The early results are encouraging: it seems that 1/ the hand position has a high influence on the key audio duration; and 2/ the hand shape has not.},
  url   = {https://aclanthology.org/2022.lrec-1.104}
}

@InProceedings{hernandezmena-EtAl:2022:LREC,
  author   = {Hernandez Mena, Carlos Daniel and Mollberg, David Erik and Borský, Michal and Guðnason, Jón},
  title   = {Samrómur Children: An Icelandic Speech Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {995--1002},
  abstract = {Samrómur Children is an Icelandic speech corpus intended for the field of automatic speech recognition. It contains 131 hours of read speech from Icelandic children aged between 4 to 17 years. The test portion was meticulously selected to cover a wide range of ages as possible; we aimed to have exactly the same amount of data per age range. The speech was collected with the crowd-sourcing platform Samrómur.is, which is inspired on the "Mozilla's Common Voice Project". The corpus was developed within the framework of the "Language Technology Programme for Icelandic 2019 - 2023"; the goal of the project is to make Icelandic available in language-technology applications. Samrómur Children is the first corpus in Icelandic with children's voices for public use under a Creative Commons license. Additionally, we present baseline experiments and results using Kaldi.},
  url   = {https://aclanthology.org/2022.lrec-1.105}
}

@InProceedings{solberg-ortiz:2022:LREC,
  author   = {Solberg, Per Erik and Ortiz, Pablo},
  title   = {The Norwegian Parliamentary Speech Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {1003--1008},
  abstract = {The Norwegian Parliamentary Speech Corpus (NPSC) is a speech dataset with recordings of meetings from Stortinget, the Norwegian parliament. It is the first, publicly available dataset containing unscripted, Norwegian speech designed for training of automatic speech recognition (ASR) systems. The recordings are manually transcribed and annotated with language codes and speakers, and there are detailed metadata about the speakers. The transcriptions exist in both normalized and non-normalized form, and non-standardized words are explicitly marked and annotated with standardized equivalents. To test the usefulness of this dataset, we have compared an ASR system trained on the NPSC with a baseline system trained on only manuscript-read speech. These systems were tested on an independent dataset containing spontaneous, dialectal speech. The NPSC-trained system performed significantly better, with a 22.9\% relative improvement in word error rate (WER). Moreover, training on the NPSC is shown to have a “democratizing” effects in terms of dialects, as improvements are generally larger for dialects with higher WER from the baseline system.},
  url   = {https://aclanthology.org/2022.lrec-1.106}
}

@InProceedings{bentum-EtAl:2022:LREC,
  author   = {Bentum, Martijn and ten Bosch, Louis and van den Heuvel, Henk and Wills, Simone and van der Niet, Domenique and Dijkstra, Jelske and Van de Velde, Hans},
  title   = {A Speech Recognizer for Frisian/Dutch Council Meetings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1009--1015},
  abstract = {We developed a bilingual Frisian/Dutch speech recognizer for council meetings in Fryslân (the Netherlands). During these meetings both Frisian and Dutch are spoken, and code switching between both languages shows up frequently. The new speech recognizer is based on an existing speech recognizer for Frisian and Dutch named FAME!, which was trained and tested on historical radio broadcasts. Adapting a speech recognizer for the council meeting domain is challenging because of acoustic background noise, speaker overlap and the jargon typically used in council meetings. To train the new recognizer, we used the radio broadcast materials utilized for the development of the FAME! recognizer and added newly created manually transcribed audio recordings of council meetings from eleven Frisian municipalities, the Frisian provincial council and the Frisian water board. The council meeting recordings consist of 49 hours of speech, with 26 hours of Frisian speech and 23 hours of Dutch speech. Furthermore, from the same sources, we obtained texts in the domain of council meetings containing 11 million words; 1.1 million Frisian words and 9.9 million Dutch words. We describe the methods used to train the new recognizer, report the observed word

error rates, and perform an error analysis on remaining errors.},
  url   = {https://aclanthology.org/2022.lrec-1.107}
}

@InProceedings{fukuda-EtAl:2022:LREC,
  author   = {Fukuda, Meiko and Nishimura, Ryota and Umezawa, Maina and Yamamoto, Kazumasa and Iribe, Yurie and Kitaoka, Norihide},
  title   = {Elderly Conversational Speech Corpus with Cognitive Impairment Test and Pilot Dementia Detection Experiment Using Acoustic Characteristics of Speech in Japanese Dialects},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1016--1022},
  abstract = {There is a need for a simple method of detecting early signs of dementia which is not burdensome to patients, since early diagnosis and treatment can often slow the advance of the disease. Several studies have explored using only the acoustic and linguistic information of conversational speech as diagnostic material, with some success. To accelerate this research, we recorded natural conversations between 128 elderly people living in four different regions of Japan and interviewers, who also administered the Hasegawa’s Dementia Scale-Revised (HDS-R), a cognitive impairment test. Using our elderly speech corpus and dementia test results, we propose an SVM-based screening method which can detect dementia using the acoustic features of conversational speech even when regional dialects are present. We accomplish this by omitting some acoustic features, to limit the negative effect of differences between dialects. When using our proposed method, a dementia detection accuracy rate of about 91\% was achieved for speakers from two regions. When speech from four regions was used in a second experiment, the discrimination rate fell to 76.6\%, but this may have been due to using only sentence-level acoustic features in the second experiment, instead of sentence and phoneme-level features as in the previous experiment. This is an on-going research project, and additional investigation is needed to understand differences in the acoustic characteristics of phoneme units in the conversational speech collected from these four regions, to determine whether the removal of formants and other features can improve the dementia detection rate.},
  url   = {https://aclanthology.org/2022.lrec-1.108}
}

@InProceedings{kocabiyikoglu-EtAl:2022:LREC,
  author   = {Kocabiyikoglu, Ali Can and Portet, François and Gibert, Prudence and Blanchon, Hervé and Babouchkine, Jean-Marc and Gavazzi, Gaëtan},
  title   = {A Spoken Drug Prescription Dataset in French for Spoken Language Understanding},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1023--1031},
  abstract = {Spoken medical dialogue systems are increasingly attracting interest to enhance access to healthcare services and improve quality and traceability of patient care. In this paper, we focus on medical drug prescriptions acquired on smartphones through spoken dialogue. Such systems would facilitate the traceability of care and would free the clinicians' time. However, there is a lack of speech corpora to develop such systems since most of the related corpora are in text form and in English. To facilitate the research and development of spoken medical dialogue systems, we present, to the best of our knowledge, the first spoken medical drug prescriptions corpus, named PxNLU. It contains 4 hours of transcribed and annotated dialogues of drug prescriptions in French acquired through an experiment with 55 participants experts and non-experts in prescriptions. We also present some experiments that demonstrate the interest of this corpus for the evaluation and development of medical dialogue systems.},
  url   = {https://aclanthology.org/2022.lrec-1.109}
}

@InProceedings{vradi-EtAl:2022:LREC,
  author   = {Váradi, Tamás and Nyéki, Bence and Koeva, Svetla and Tadić, Marko and Štefanec, Vanja and Ogrodniczuk, Maciej and Nitoń, Bartłomiej and Pęzik, Piotr and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Tufiș, Dan and Garabík, Radovan and Krek, Simon and Repar, Andraž},
  title   = {Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {100--108},
  abstract = {This article presents the current outcomes of the CURLICAT CEF Telecom project, which aims to collect and deeply annotate a set of large corpora from selected domains. The CURLICAT corpus includes 7 monolingual corpora (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) containing selected samples from respective national corpora. These corpora are automatically tokenized, lemmatized and morphologically analysed and the named entities annotated. The annotations are uniformly provided for each language specific corpus while the common metadata schema is harmonised across the languages. Additionally, the corpora are annotated for IATE terms in all languages. The file format is CoNLL-U Plus format, containing the ten columns specific to the CoNLL-U format and three extra columns specific to our corpora as defined by Varádi et al. (2020). The CURLICAT corpora represent a rich and

valuable source not just for training NMT models, but also for further studies and developments in machine learning, cross-lingual terminological data extraction and classification.},
  url   = {https://aclanthology.org/2022.lrec-1.11}
}

@InProceedings{tejedorgarca-EtAl:2022:LREC,
  author   = {Tejedor-García, Cristian and van der Molen, Berrie and van den Heuvel, Henk and van Hessen, Arjan and Pieters, Toine},
  title   = {Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1032--1039},
  abstract = {The current largest open-source generic automatic speech recognition (ASR) system for Dutch, Kaldi_NL, does not include a domain-specific healthcare jargon in the lexicon. Commercial alternatives (e.g., Google ASR system) are also not suitable for this purpose, not only because of the lexicon issue, but they do not safeguard privacy of sensitive data sufficiently and reliably. These reasons motivate that just a small amount of medical staff employs speech technology in the Netherlands. This paper proposes an innovative ASR training method developed within the Homo Medicinalis (HoMed) project. On the semantic level it specifically targets automatic transcription of doctor-patient consultation recordings with a focus on the use of medicines. In the first stage of HoMed, the Kaldi_NL language model (LM) is fine-tuned with lists of Dutch medical terms and transcriptions of Dutch online healthcare news bulletins. Despite the acoustic challenges and linguistic complexity of the domain, we reduced the word error rate (WER) by 5.2\%. The proposed method could be employed for ASR domain adaptation to other domains with sensitive and special category data. These promising results allow us to apply this methodology on highly sensitive audiovisual recordings of patient consultations at the Netherlands Institute for Health Services Research (Nivel).},
  url   = {https://aclanthology.org/2022.lrec-1.110}
}

@InProceedings{moutti-EtAl:2022:LREC,
  author   = {Moutti, Maria and Eleftheriou, Sofia and Koromilas, Panagiotis and Giannakopoulos, Theodoros},
  title   = {A Dataset for Speech Emotion Recognition in Greek Theatrical Plays},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {1040--1046},
  abstract = {Machine learning methodologies can be adopted in cultural applications and propose new ways to distribute or even present the cultural content to the public. For instance, speech analytics can be adopted to automatically generate subtitles in theatrical plays, in order to (among other purposes) help people with hearing loss. Apart from a typical speech-to-text transcription with Automatic Speech Recognition (ASR), Speech Emotion Recognition (SER) can be used to automatically predict the underlying emotional content of speech dialogues in theatrical plays, and thus to provide a deeper understanding how the actors utter their lines. However, real-world datasets from theatrical plays are not available in the literature. In this work we present GreThE, the Greek Theatrical Emotion dataset, a new publicly available data collection for speech emotion recognition in Greek theatrical plays. The dataset contains utterances from various actors and plays, along with respective valence and arousal annotations. Towards this end, multiple annotators have been asked to provide their input for each speech recording and inter-annotator agreement is taken into account in the final ground truth generation. In addition, we discuss the results of some indicative experiments that have been conducted with machine and deep learning frameworks, using the dataset, along with some widely used databases in the field of speech emotion recognition.},
  url   = {https://aclanthology.org/2022.lrec-1.111}
}

@InProceedings{piits-EtAl:2022:LREC,
  author   = {Piits, Liisi and Pajupuu, Hille and Sahkai, Heete and Altrov, Rene and Ermus, Liis and Tamuri, Kairi and Hein, Indrek and Mihkla, Meelis and Kiissel, Indrek and Männisalu, Egert and Suluste, Kristjan and Pajupuu, Jaan},
  title   = {Audiobook Dialogues as Training Data for Conversational Style Synthetic Voices},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1047--1053},
  abstract = {Synthetic voices are increasingly used in applications that require a conversational speaking style, raising the question as to which type of training data yields the most suitable speaking style for such applications. This study compares voices trained on three corpora of equal size recorded by the same speaker: an audiobook character speech (dialogue) corpus, an audiobook narrator speech corpus, and a neutral-style sentence-based corpus. The voices were trained with three text-to-speech synthesisers: two hidden Markov model-based synthesisers and a neural synthesiser. An evaluation study tested the suitability of their speaking style for use in customer service voice chatbots. Independently of the synthesiser used, the voices trained on the character speech corpus received the lowest, and those trained on the neutral-style corpus the highest scores. However, the evaluation

results may have been confounded by the greater acoustic variability, less balanced sentence length distribution, and poorer phonemic coverage of the character speech corpus, especially compared to the neutral-style corpus. Therefore, the next step will be the creation of a more uniform, balanced, and representative audiobook dialogue corpus, and the evaluation of its suitability for further conversational-style applications besides customer service chatbots.},
  url   = {https://aclanthology.org/2022.lrec-1.112}
}

@InProceedings{wu-EtAl:2022:LREC1,
  author   = {WU, Yaru and Suchanek, Fabian and Vasilescu, Ioana and Lamel, Lori and Adda-Decker, Martine},
  title   = {Using a Knowledge Base to Automatically Annotate Speech Corpora and to Identify Sociolinguistic Variation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1054--1060},
  abstract = {Speech characteristics vary from speaker to speaker. While some variation phenomena are due to the overall communication setting, others are due to diastratic factors such as gender, provenance, age, and social background. The analysis of these factors, although relevant for both linguistic and speech technology communities, is hampered by the need to annotate existing corpora or to recruit, categorise, and record volunteers as a function of targeted profiles. This paper presents a methodology that uses a knowledge base to provide speaker-specific information. This can facilitate the enrichment of existing corpora with new annotations extracted from the knowledge base. The method also helps the large scale analysis by automatically extracting instances of speech variation to correlate with diastratic features. We apply our method to an over 120-hour corpus of broadcast speech in French and investigate variation patterns linked to reduction phenomena and/or specific to connected speech such as disfluencies. We find significant differences in speech rate, the use of filler words, and the rate of non-canonical realisations of frequent segments as a function of different professional categories and age groups.},
  url   = {https://aclanthology.org/2022.lrec-1.113}
}

@InProceedings{li-EtAl:2022:LREC1,
  author   = {Li, Xinjian and Metze, Florian and Mortensen, David R. and Black, Alan W and Watanabe, Shinji},
  title   = {Phone Inventories and Recognition for Every Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {1061--1067},
abstract = {Identifying phone inventories is a crucial component in language documentation and the preservation of endangered languages. However, even the largest collection of phone inventory only covers about 2000 languages, which is only 1/4 of the total number of languages in the world. A majority of the remaining languages are endangered. In this work, we attempt to solve this problem by estimating the phone inventory for any language listed in Glottolog, which contains phylogenetic information regarding 8000 languages. In particular, we propose one probabilistic model and one non-probabilistic model, both using phylogenetic trees (``language family trees'') to measure the distance between languages. We show that our best model outperforms baseline models by 6.5 F1. Furthermore, we demonstrate that, with the proposed inventories, the phone recognition model can be customized for every language in the set, which improved the PER (phone error rate) in phone recognition by 25\%.},
url = {https://aclanthology.org/2022.lrec-1.114}
}

@InProceedings{roussis-EtAl:2022:LREC1,
author = {Roussis, Dimitrios and Papavassiliou, Vassilis and Sofianopoulos, Sokratis and Prokopidis, Prokopis and Piperidis, Stelios},
title = {Constructing Parallel Corpora from COVID-19 News using MediSys Metadata},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {1068--1072},
abstract = {This paper presents a collection of parallel corpora generated by exploiting the COVID-19 related dataset of metadata created with the Europe Media Monitor (EMM) / Medical Information System (MediSys) processing chain of news articles. We describe how we constructed comparable monolingual corpora of news articles related to the current pandemic and used them to mine about 11.2 million segment alignments in 26 EN-X language pairs, covering most official EU languages plus Albanian, Arabic, Icelandic, Macedonian, and Norwegian. Subsets of this collection have been used in shared tasks (e.g. Multilingual Semantic Search, Machine Translation) aimed at accelerating the creation of resources and tools needed to facilitate access to information in the COVID-19 emergency situation.},
url = {https://aclanthology.org/2022.lrec-1.115}
}

@InProceedings{zhang-EtAl:2022:LREC,
author = {Zhang, Dongxu and Mohan, Sunil and Torkar, Michaela and McCallum, Andrew},

  title   = {A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1073--1082},
  abstract = {We introduce ChemDisGene, a new dataset for training and evaluating multi-class multi-label biomedical relation extraction models. Our dataset contains 80k biomedical research abstracts labeled with mentions of chemicals, diseases, and genes, portions of which human experts labeled with 18 types of biomedical relationships between these entities (intended for evaluation), and the remainder of which (intended for training) has been distantly labeled via the CTD database with approximately 78\% accuracy. In comparison to similar preexisting datasets, ours is both substantially larger and cleaner; it also includes annotations linking mentions to their entities. We also provide three baseline deep neural network relation extraction models trained and evaluated on our new dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.116}
}

@InProceedings{bardhan-EtAl:2022:LREC,
  author   = {Bardhan, Jayetri and Colas, Anthony and Roberts, Kirk and Wang, Daisy Zhe},
  title   = {DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records For Medicine Related Queries},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1083--1097},
  abstract = {This paper develops the first question answering dataset (DrugEHRQA) containing question-answer pairs from both structured tables and unstructured notes from a publicly available Electronic Health Record (EHR). EHRs contain patient records, stored in structured tables and unstructured clinical notes. The information in structured and unstructured EHRs is not strictly disjoint: information may be duplicated, contradictory, or provide additional context between these sources. Our dataset has medication-related queries, containing over 70,000 question-answer pairs. To provide a baseline model and help analyze the dataset, we have used a simple model (MultimodalEHRQA) which uses the predictions of a modality selection network to choose between EHR tables and clinical notes to answer the questions. This is used to direct the questions to the table-based or text-based state-of-the-art QA model. In order to address the problem arising from complex, nested queries, this is the first time Relation-Aware Schema

Encoding and Linking for Text-to-SQL Parsers (RAT-SQL) has been used to test the structure of query templates in EHR data. Our goal is to provide a benchmark dataset for multi-modal QA systems, and to open up new avenues of research in improving question answering over EHR structured data by using context from unstructured clinical data.},
  url   = {https://aclanthology.org/2022.lrec-1.117}
}

@InProceedings{verkijk-vossen:2022:LREC,
  author   = {Verkijk, Stella and Vossen, Piek},
  title   = {Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1098--1103},
  abstract = {Neural Network (NN) architectures are used more and more to model large amounts of data, such as text data available online. Transformer-based NN architectures have shown to be very useful for language modelling. Although many researchers study how such Language Models (LMs) work, not much attention has been paid to the privacy risks of training LMs on large amounts of data and publishing them online. This paper presents a new method for anonymizing a language model by presenting the way in which MedRoBERTa.nl, a Dutch language model for hospital notes, was anonymized. The two-step method involves i) automatic anonymization of the training data and ii) semi-automatic anonymization of the LM’s vocabulary. Adopting the fill-mask task where the model predicts what tokens are most probable in a certain context, it was tested how often the model will predict a name in a context where a name should be. It was shown that it predicts a name-like token 0.2\% of the time. Any name-like token that was predicted was never the name originally present in the training data. By explaining how a LM trained on highly private real-world medical data can be published, we hope that more language resources will be published openly and responsibly so the scientific community can profit from them.},
  url   = {https://aclanthology.org/2022.lrec-1.118}
}

@InProceedings{grobol-EtAl:2022:LREC,
  author   = {Grobol, Loïc and Regnault, Mathilde and Ortiz Suarez, Pedro and Sagot, Benoît and Romary, Laurent and Crabbé, Benoit},
  title   = {BERTrade: Using Contextual Embeddings to Parse Old French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {1104--1113},
  abstract = {The successes of contextual word embeddings learned by training large-scale language models, while remarkable, have mostly occurred for languages where significant amounts of raw texts are available and where annotated data in downstream tasks have a relatively regular spelling. Conversely, it is not yet completely clear if these models are also well suited for lesser-resourced and more irregular languages. We study the case of Old French, which is in the interesting position of having relatively limited amount of available raw text, but enough annotated resources to assess the relevance of contextual word embedding models for downstream NLP tasks. In particular, we use POS-tagging and dependency parsing to evaluate the quality of such models in a large array of configurations, including models trained from scratch from small amounts of raw text and models pre-trained on other languages but fine-tuned on Medieval French data.},
  url   = {https://aclanthology.org/2022.lrec-1.119}
}

@InProceedings{rytting-EtAl:2022:LREC,
  author   = {Rytting, C. Anton and Novak, Valerie and Hull, James R. and Frank, Victor M. and Rodrigues, Paul and Lee, Jarrett G. W. and Miller-Sims, Laurel},
  title   = {RU-ADEPT: Russian Anonymized Dataset with Eight Personality Traits},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {109--118},
  abstract = {Social media has provided a platform for many individuals to easily express themselves naturally and publicly, and researchers have had the opportunity to utilize large quantities of this data to improve author trait analysis techniques and to improve author trait profiling systems. The majority of the work in this area, however, has been narrowly spent on English and other Western European languages, and generally focuses on a single social network at a time, despite the large quantity of data now available across languages and differences that have been found across platforms. This paper introduces RU-ADEPT, a dataset of Russian authors' personality trait scores--Big Five and Dark Triad, demographic information (e.g. age, gender), with associated corpus of the authors' cross-contributions to (up to) four different social media platforms--VKontakte (VK), LiveJournal, Blogger, and Moi Mir. We believe this to be the first publicly-available dataset associating demographic and personality trait data with Russian-language social media content, the first paper to describe the collection of Dark Triad scores with texts across multiple Russian-language social media platforms, and to a limited extent, the first publicly-available dataset of personality traits to author content across several different social media sites.},

  url   = {https://aclanthology.org/2022.lrec-1.12}
}

@InProceedings{kanerva-ginter:2022:LREC,
  author   = {Kanerva, Jenna and Ginter, Filip},
  title   = {Out-of-Domain Evaluation of Finnish Dependency
Parsing},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1114--1124},
  abstract = {The prevailing practice in the academia is to
evaluate the model performance on in-domain evaluation data
typically set aside from the training corpus. However, in many real
world applications the data on which the model is applied may very
substantially differ from the characteristics of the training data.
In this paper, we focus on Finnish out-of-domain parsing by
introducing a novel UD Finnish-OOD out-of-domain treebank including
five very distinct data sources (web documents, clinical, online
discussions, tweets, and poetry), and a total of 19,382 syntactic
words in 2,122 sentences released under the Universal Dependencies
framework. Together with the new treebank, we present extensive out-
of-domain parsing evaluation utilizing the available section-level
information from three different Finnish UD treebanks (TDT, PUD,
OOD). Compared to the previously existing treebanks, the new
Finnish-OOD is shown include sections more challenging for the
general parser, creating an interesting evaluation setting and
yielding valuable information for those applying the parser outside
of its training domain.},
  url   = {https://aclanthology.org/2022.lrec-1.120}
}

@InProceedings{gugliotta-dinarelli:2022:LREC,
  author   = {gugliotta, elisa and Dinarelli, Marco},
  title   = {TArC: Tunisian Arabish Corpus, First complete
release},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1125--1136},
  abstract = {In this paper we present the final result of a
project focused on Tunisian Arabic encoded in Arabizi, the Latin-
based writing system for digital conversations. The project led to
the realization of two integrated and independent tools: a
linguistic corpus and a neural network architecture created to
annotate the former with various levels of linguistic information
(code-switching classification, transliteration, tokenization, POS-
tagging, lemmatization). We discuss the choices made in terms of

computational and linguistic methodology and the strategies adopted to improve our results. We report on the experiments performed in order to outline our research path. Finally, we explain the reasons why we believe in the potential of these tools for both computational and linguistic researches.},
  url   = {https://aclanthology.org/2022.lrec-1.121}
}

@InProceedings{abokrtsk-EtAl:2022:LREC,
  author   = {Žabokrtský, Zdeněk and Bafna, Niyati and Bodnár, Jan and Kyjánek, Lukáš and Svoboda, Emil and Ševčíková, Magda and Vidra, Jonáš},
  title   = {Towards Universal Segmentations: UniSegments 1.0},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1137--1149},
  abstract = {Our work aims at developing a multilingual data resource for morphological segmentation. We present a survey of 17 existing data resources relevant for segmentation in 32 languages, and analyze diversity of how individual linguistic phenomena are captured across them. Inspired by the success of Universal Dependencies, we propose a harmonized scheme for segmentation representation, and convert the data from the studied resources into this common scheme. Harmonized versions of resources available under free licenses are published as a collection called UniSegments 1.0.},
  url   = {https://aclanthology.org/2022.lrec-1.122}
}

@InProceedings{moran-EtAl:2022:LREC,
  author   = {Moran, Steven and Bentz, Christian and Gutierrez-Vasques, Ximena and Sozinova, Olga and Samardzic, Tanja},
  title   = {TeDDi Sample: Text Data Diversity Sample for Language Comparison and Multilingual NLP},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1150--1158},
  abstract = {We present the TeDDi sample, a diversity sample of text data for language comparison and multilingual Natural Language Processing. The TeDDi sample currently features 89 languages based on the typological diversity sample in the World Atlas of Language Structures. It consists of more than 20k texts and is accompanied by open-source corpus processing tools. The aim of TeDDi is to facilitate text-based quantitative analysis of linguistic diversity. We describe in detail the TeDDi sample, how it was created, data availability, and its added value through for NLP and linguistic

research.},
  url   = {https://aclanthology.org/2022.lrec-1.123}
}

@InProceedings{bear-cook:2022:LREC,
  author   = {Bear, Diego and Cook, Paul},
  title   = {Leveraging a Bilingual Dictionary to Learn Wolastoqey Word Representations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1159--1166},
  abstract = {Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) have been used to bolster the performance of natural language processing systems in a wide variety of tasks, including information retrieval (Roy et al., 2018) and machine translation (Qi et al., 2018). However, approaches to learning word embeddings typically require large corpora of running text to learn high quality representations. For many languages, such resources are unavailable. This is the case for Wolastoqey, also known as Passamaquoddy-Maliseet, an endangered low-resource Indigenous language. As there exist no large corpora of running text for Wolastoqey, in this paper, we leverage a bilingual dictionary to learn Wolastoqey word embeddings by encoding their corresponding English definitions into vector representations using pretrained English word and sequence representation models. Specifically, we consider representations based on pretrained word2vec (Mikolov et al., 2013), RoBERTa (Liu et al., 2019) and sentence-BERT (Reimers and Gurevych, 2019) models. We evaluate these embeddings in word prediction tasks focused on part-of-speech, animacy, and transitivity; semantic clustering; and reverse dictionary search. In all evaluations we demonstrate that approaches using these embeddings outperform task-specific baselines, without requiring any language-specific training or fine-tuning.},
  url   = {https://aclanthology.org/2022.lrec-1.124}
}

@InProceedings{wiechetek-EtAl:2022:LREC,
  author   = {Wiechetek, Linda and Hiovain-Asikainen, Katri and Mikkelsen, Inga Lill Sigga and Moshagen, Sjur and Pirinen, Flammie and Trosterud, Trond and Gaup, Børre},
  title   = {Unmasking the Myth of Effortless Big Data - Making an Open Source Multi-lingual Infrastructure and Building Language Resources from Scratch},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1167--1177},

  abstract = {Machine learning (ML) approaches have dominated NLP during the last two decades. From machine translation and speech technology, ML tools are now also in use for spellchecking and grammar checking, with a blurry distinction between the two. We unmask the myth of effortless big data by illuminating the efforts and time that lay behind building a multi-purpose corpus with regard to collecting, mark-up and building from scratch. We also discuss what kind of language technology minority languages actually need, and to what extent the dominating paradigm has been able to deliver these tools. In this context we present our alternative to corpus-based language technology, which is knowledge-based language technology, and we show how this approach can provide language technology solutions for languages being outside the reach of machine learning procedures. We present a stable and mature infrastructure (GiellaLT) containing more than hundred languages and building a number of language technology tools that are useful for language communities.},
  url   = {https://aclanthology.org/2022.lrec-1.125}
}

@InProceedings{liesenfeld-dingemanse:2022:LREC,
  author   = {Liesenfeld, Andreas and Dingemanse, Mark},
  title   = {Building and curating conversational corpora for diversity-aware language science and technology},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1178--1192},
  abstract = {We present an analysis pipeline and best practice guidelines for building and curating corpora of everyday conversation in diverse languages. Surveying language documentation corpora and other resources that cover 67 languages and varieties from 28 phyla, we describe the compilation and curation process, specify minimal properties of a unified format for interactional data, and develop methods for quality control that take into account turn-taking and timing. Two case studies show the broad utility of conversational data for (i) charting human interactional infrastructure and (ii) tracing challenges and opportunities for current ASR solutions. Linguistically diverse conversational corpora can provide new insights for the language sciences and stronger empirical foundations for language technology.},
  url   = {https://aclanthology.org/2022.lrec-1.126}
}

@InProceedings{przybyl-EtAl:2022:LREC,
  author   = {Przybyl, Heike and Lapshinova-Koltunski, Ekaterina and Menzel, Katrin and Fischer, Stefan and Teich, Elke},
  title   = {EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1193--1200},
  abstract = {In this paper, we describe the creation and annotation of EPIC UdS, a multilingual corpus of simultaneous interpreting for English, German and Spanish. We give an overview of the comparable and parallel, aligned corpus variants and explore various applications of the corpus. What makes EPIC UdS relevant is that it is one of the rare interpreting corpora that includes transcripts suitable for research on more than one language pair and on interpreting with regard to German. It not only contains transcribed speeches, but also rich metadata and fine-grained linguistic annotations tailored for diverse applications across a broad range of linguistic subfields.},
  url   = {https://aclanthology.org/2022.lrec-1.127}
}

@InProceedings{green-maynard-lin:2022:LREC,
  author   = {Green, Thomas and Maynard, Diana and Lin, Chenghua},
  title   = {Development of a Benchmark Corpus to Support Entity Recognition in Job Descriptions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1201--1208},
  abstract = {We present the development of a benchmark suite consisting of an annotation schema, training corpus and baseline model for Entity Recognition (ER) in job descriptions, published under a Creative Commons license. This was created to address the distinct lack of resources available to the community for the extraction of salient entities, such as skills, from job descriptions. The dataset contains 18.6k entities comprising five types (Skill, Qualification, Experience, Occupation, and Domain). We include a benchmark CRF-based ER model which achieves an F1 score of 0.59. Through the establishment of a standard definition of entities and training/testing corpus, the suite is designed as a foundation for future work on tasks such as the development of job recommender systems.},
  url   = {https://aclanthology.org/2022.lrec-1.128}
}

@InProceedings{arrigo-EtAl:2022:LREC,
  author   = {Arrigo, Michael and Strassel, Stephanie and King, Nolan and Tran, Thao and Mason, Lisa},
  title   = {CAMIO: A Corpus for OCR in Multiple Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1209--1216},
  abstract = {CAMIO (Corpus of Annotated Multilingual Images for OCR) is a new corpus created by Linguistic Data Consortium to serve as a resource to support the development and evaluation of optical character recognition (OCR) and related technologies for 35 languages across 24 unique scripts. The corpus comprises nearly 70,000 images of machine printed text, covering a wide variety of topics and styles, document domains, attributes and scanning/capture artifacts. Most images have been exhaustively annotated for text localization, resulting in over 2.3M line-level bounding boxes. For 13 of the 35 languages, 1250 images/language have been further annotated with orthographic transcriptions of each line plus specification of reading order, yielding over 2.4M tokens of transcribed text. The resulting annotations are represented in a comprehensive XML output format defined for this corpus. The paper discusses corpus design and implementation, challenges encountered, baseline performance results obtained on the corpus for text localization and OCR decoding, and plans for corpus publication.},
  url   = {https://aclanthology.org/2022.lrec-1.129}
}

@InProceedings{brabant-lecorv-rojasbarahona:2022:LREC,
  author   = {Brabant, Quentin and Lecorvé, Gwénolé and Rojas Barahona, Lina M.},
  title   = {CoQAR: Question Rewriting on CoQA},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {119--126},
  abstract = {Questions asked by humans during a conversation often contain contextual dependencies, i.e., explicit or implicit references to previous dialogue turns. These dependencies take the form of coreferences (e.g., via pronoun use) or ellipses, and can make the understanding difficult for automated systems. One way to facilitate the understanding and subsequent treatments of a question is to rewrite it into an out-of-context form, i.e., a form that can be understood without the conversational context. We propose CoQAR, a corpus containing 4.5K conversations from the Conversational Question-Answering dataset CoQA, for a total of 53K follow-up question-answer pairs. Each original question was manually annotated with at least 2 at most 3 out-of-context rewritings. CoQA originally contains 8k conversations, which sum up to 127k question-answer pairs. CoQAR can be used in the supervised learning of three tasks: question paraphrasing, question rewriting and conversational question answering. In order to assess the quality of CoQAR's rewritings, we conduct several experiments consisting in training and evaluating models for these three tasks. Our results support the idea that question rewriting can be used as a preprocessing step for

(conversational and non-conversational) question answering models, thereby increasing their performances.},
  url   = {https://aclanthology.org/2022.lrec-1.13}
}

@InProceedings{wilkens-EtAl:2022:LREC,
  author   = {Wilkens, Rodrigo and Alfter, David and Wang, Xiaoou and Pintard, Alice and Tack, Anaïs and Yancey, Kevin P. and François, Thomas},
  title   = {FABRA: French Aggregator-Based Readability Assessment toolkit},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1217--1233},
  abstract = {In this paper, we present the FABRA: readability toolkit based on the aggregation of a large number of readability predictor variables. The toolkit is implemented as a service-oriented architecture, which obviates the need for installation, and simplifies its integration into other projects. We also perform a set of experiments to show which features are most predictive on two different corpora, and how the use of aggregators improves performance over standard feature-based readability prediction. Our experiments show that, for the explored corpora, the most important predictors for native texts are measures of lexical diversity, dependency counts and text coherence, while the most important predictors for foreign texts are syntactic variables illustrating language development, as well as features linked to lexical sophistication. FABRA: have the potential to support new research on readability assessment for French.},
  url   = {https://aclanthology.org/2022.lrec-1.130}
}

@InProceedings{aicher-EtAl:2022:LREC2,
  author   = {Aicher, Annalena and Gerstenlauer, Nadine and Feustel, Isabel and Minker, Wolfgang and Ultes, Stefan},
  title   = {Towards Building a Spoken Dialogue System for Argument Exploration},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1234--1241},
  abstract = {Speech interfaces for argumentative dialogue systems (ADS) are rather scarce. The complex task they pursue hinders the application of common natural language understanding (NLU) approaches in this domain. To address this issue we include an adaption of a recently introduced NLU framework tailored to argumentative tasks into a complete ADS. We evaluate the likeability

and motivation of users to interact with the new system in a user study. Therefore, we compare it to a solid baseline utilizing a drop-down menu. The results indicate that the integration of a flexible NLU framework enables a far more natural and satisfying interaction with human users in real-time. Even though the drop-down menu convinces regarding its robustness, the willingness to use the new system is significantly higher. Hence, the featured NLU framework provides a sound basis to build an intuitive interface which can be extended to adapt its behavior to the individual user.},
  url   = {https://aclanthology.org/2022.lrec-1.131}
}

@InProceedings{park-EtAl:2022:LREC2,
  author   = {park, chanjun and Jang, Yoonna and Lee, Seolhwa and Park, Sungjin and Lim, Heuiseok},
  title   = {FreeTalky: Don’t Be Afraid! Conversations Made Easier by a Humanoid Robot using Persona-based Dialogue},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1242--1248},
  abstract = {We propose a deep learning-based foreign language learning platform, named FreeTalky, for people who experience anxiety dealing with foreign languages, by employing a humanoid robot NAO and various deep learning models. A persona-based dialogue system that is embedded in NAO provides an interesting and consistent multi-turn dialogue for users. Also, an grammar error correction system promotes improvement in grammar skills of the users. Thus, our system enables personalized learning based on persona dialogue and facilitates grammar learning of a user using grammar error feedback. Furthermore, we verified whether FreeTalky provides practical help in alleviating xenoglossophobia by replacing the real human in the conversation with a NAO robot, through human evaluation.},
  url   = {https://aclanthology.org/2022.lrec-1.132}
}

@InProceedings{hayashibe:2022:LREC,
  author   = {Hayashibe, Yuta},
  title   = {Self-Contained Utterance Description Corpus for Japanese Dialog},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1249--1255},
  abstract = {Often both an utterance and its context must be read to understand its intent in a dialog. Herein we propose a task,

Self- Contained Utterance Description (SCUD), to describe the intent of an utterance in a dialog with multiple simple natural sentences without the context. If a task can be performed concurrently with high accuracy as the conversation continues such as in an accommodation search dialog, the operator can easily suggest candidates to the customer by inputting SCUDs of the customer's utterances to the accommodation search system. SCUDs can also describe the transition of customer requests from the dialog log. We construct a Japanese corpus to train and evaluate automatic SCUD generation. The corpus consists of 210 dialogs containing 10,814 sentences. We conduct an experiment to verify that SCUDs can be automatically generated. Additionally, we investigate the influence of the amount of training data on the automatic generation performance using 8,200 additional examples.},
  url   = {https://aclanthology.org/2022.lrec-1.133}
}

@InProceedings{huynh-EtAl:2022:LREC,
  author   = {Huynh, Jessica and Chiang, Ting-Rui and Bigham, Jeffrey and Eskenazi, Maxine},
  title   = {DialCrowd 2.0: A Quality-Focused Dialog System Crowdsourcing Toolkit},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1256--1263},
  abstract = {Dialog system developers need high-quality data to train, fine-tune and assess their systems. They often use crowdsourcing for this since it provides large quantities of data from many workers. However, the data may not be of sufficiently good quality. This can be due to the way that the requester presents a task and how they interact with the workers. This paper introduces DialCrowd 2.0 to help requesters obtain higher quality data by, for example, presenting tasks more clearly and facilitating effective communication with workers. DialCrowd 2.0 guides developers in creating improved Human Intelligence Tasks (HITs) and is directly applicable to the workflows used currently by developers and researchers.},
  url   = {https://aclanthology.org/2022.lrec-1.134}
}

@InProceedings{gonalooliveira-EtAl:2022:LREC,
  author   = {Gonçalo Oliveira, Hugo and Ferreira, Patrícia and Martins, Daniel and Silva, Catarina and Alves, Ana},
  title   = {A Brief Survey of Textual Dialogue Corpora},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {1264--1274},
  abstract = {Several dialogue corpora are currently available for research purposes, but they still fall short for the growing interest in the development of dialogue systems with their own specific requirements. In order to help those requiring such a corpus, this paper surveys a range of available options, in terms of aspects like speakers, size, languages, collection, annotations, and domains. Some trends are identified and possible approaches for the creation of new corpora are also discussed.},
  url   = {https://aclanthology.org/2022.lrec-1.135}
}

@InProceedings{rckert-EtAl:2022:LREC,
  author   = {Rückert, Ulrich and Sunkara, Srinivas and Rastogi, Abhinav and Prakash, Sushant and Khaitan, Pranav},
  title   = {A Unified Approach to Entity-Centric Context Tracking in Social Conversations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1275--1285},
  abstract = {In human-human conversations, Context Tracking deals with identifying important entities and keeping track of their properties and relationships. This is a challenging problem that encompasses several subtasks such as slot tagging, coreference resolution, resolving plural mentions and entity linking. We approach this problem as an end-to-end modeling task where the conversational context is represented by an entity repository containing the entity references mentioned so far, their properties and the relationships between them. The repository is updated turn-by-turn, thus making training and inference computationally efficient even for long conversations. This paper lays the groundwork for an investigation of this framework in two ways. First, we release Contrack, a large scale human-human conversation corpus for context tracking with people and location annotations. It contains over 7000 conversations with an average of 11.8 turns, 5.8 entities and 15.2 references per conversation. Second, we open-source a neural network architecture for context tracking. Finally we compare this network to state-of-the-art approaches for the subtasks it subsumes and report results on the involved tradeoffs.},
  url   = {https://aclanthology.org/2022.lrec-1.136}
}

@InProceedings{hudeek-EtAl:2022:LREC,
  author   = {Hudeček, Vojtěch and Schaub, leon-paul and Stancl, Daniel and Paroubek, Patrick and Dušek, Ondřej},
  title   = {A Unifying View On Task-oriented Dialogue Annotation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1286--1296},
  abstract = {Every model is only as strong as the data that it is trained on. In this paper, we present a new dataset, obtained by merging four publicly available annotated corpora for task-oriented dialogues in several domains (MultiWOZ 2.2, CamRest676, DSTC2 and Schema-Guided Dialogue Dataset). This way, we assess the feasibility of providing a unified ontology and annotation schema covering several domains with a relatively limited effort. We analyze the characteristics of the resulting dataset along three main dimensions: language, information content and performance. We focus on aspects likely to be pertinent for improving dialogue success, e.g. dialogue consistency. Furthermore, to assess the usability of this new corpus, we thoroughly evaluate dialogue generation performance under various conditions with the help of two prominent recent end-to-end dialogue models: MarCo and GPT-2. These models were selected as popular open implementations representative of the two main dimensions of dialogue modelling. While we did not observe a significant gain for dialogue state tracking performance, we show that using more training data from different sources can improve language modelling capabilities and positively impact dialogue flow (consistency). In addition, we provide the community with one of the largest open dataset for machine learning experiments.},
  url   = {https://aclanthology.org/2022.lrec-1.137}
}

@InProceedings{origlia-EtAl:2022:LREC,
  author   = {Origlia, Antonio and Di Bratto, Martina and Di Maro, Maria and Mennella, Sabrina},
  title   = {A Multi-source Graph Representation of the Movie Domain for Recommendation Dialogues Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1297--1306},
  abstract = {In dialogue analysis, characterising named entities in the domain of interest is relevant in order to understand how people are making use of them for argumentation purposes. The movie recommendation domain is a frequently considered case study for many applications and by linguistic studies and, since many different resources have been collected throughout the years to describe it, a single database combining all these data sources is a valuable asset for cross-disciplinary investigations. We propose an integrated graph-based structure of multiple resources, enriched with the results of the application of graph analytics approaches to provide an encompassing view of the domain and of the way people talk about it during the recommendation task. While we cannot distribute the final resource because of licensing issues, we share the code to assemble and process it once the reference data have been obtained

from the original sources.},
  url   = {https://aclanthology.org/2022.lrec-1.138}
}

@InProceedings{plazadelarco-EtAl:2022:LREC,
  author   = {Plaza-del-Arco, Flor Miriam and Parras Portillo, Ana Belén and López Úbeda, Pilar and Gil, Beatriz and Martín-Valdivia, María-Teresa},
  title   = {SHARE: A Lexicon of Harmful Expressions by Spanish Speakers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1307--1316},
  abstract = {In this paper we present SHARE, a new lexical resource with 10,125 offensive terms and expressions collected from Spanish speakers. We retrieve this vocabulary using an existing chatbot developed to engage a conversation with users and collect insults via Telegram, named Fiero. This vocabulary has been manually labeled by five annotators obtaining a kappa coefficient agreement of 78.8\%. In addition, we leverage the lexicon to release the first corpus in Spanish for offensive span identification research named OffendES_spans. Finally, we show the utility of our resource as an interpretability tool to explain why a comment may be considered offensive.},
  url   = {https://aclanthology.org/2022.lrec-1.139}
}

@InProceedings{aicher-EtAl:2022:LREC1,
  author   = {Aicher, Annalena and Gerstenlauer, Nadine and Minker, Wolfgang and Ultes, Stefan},
  title   = {User Interest Modelling in Argumentative Dialogue Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {127--136},
  abstract = {Most systems helping to provide structured information and support opinion building, discuss with users without considering their individual interest. The scarce existing research on user interest in dialogue systems depends on explicit user feedback. Such systems require user responses that are not content-related and thus, tend to disturb the dialogue flow. In this paper, we present a novel model for implicitly estimating user interest during argumentative dialogues based on semantically clustered data. Therefore, an online user study was conducted to acquire training data which was used to train a binary neural network classifier in order to predict whether or not users are still interested in the

content of the ongoing dialogue. We achieved a classification accuracy of 74.9\% and furthermore investigated with different Artificial Neural Networks (ANN) which new argument would fit the user interest best.},
  url   = {https://aclanthology.org/2022.lrec-1.14}
}

@InProceedings{ylonen:2022:LREC,
  author   = {Ylonen, Tatu},
  title   = {Wiktextract: Wiktionary as Machine-Readable Structured Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1317--1325},
  abstract = {We present a machine-readable structured data version of Wiktionary. Unlike previous Wiktionary extractions, the new extractor, Wiktextract, fully interprets and expands templates and Lua modules in Wiktionary. This enables it to perform a more complete, robust, and maintainable extraction. The extracted data is multilingual and includes lemmas, inflected forms, translations, etymology, usage examples, pronunciations (including URLs of sound files), lexical and semantic relations, and various morphological, syntactic, semantic, topical, and dialectal annotations. We extract all data from the English Wiktionary. Comparing against previous extractions from language-specific dictionaries, we find that its coverage for non-English languages often matches or exceeds the coverage in the language-specific editions, with the added benefit that all glosses are in English. The data is freely available and regularly updated, enabling anyone to add more data and correct errors by editing Wiktionary. The extracted data is in JSON format and designed to be easy to use by researchers, downstream resources, and application developers.},
  url   = {https://aclanthology.org/2022.lrec-1.140}
}

@InProceedings{holmer-rennes:2022:LREC,
  author   = {Holmer, Daniel and Rennes, Evelina},
  title   = {NyLLex: A Novel Resource of Swedish Words Annotated with Reading Proficiency Level},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1326--1331},
  abstract = {What makes a text easy to read or not, depends on a variety of factors. One of the most prominent is, however, if the text contains easy, and avoids difficult, words. Deciding if a word is easy or difficult is not a trivial task, since it depends on

characteristics of the word in itself as well as the reader, but it can be facilitated by the help of a corpus annotated with word frequencies and reading proficiency levels. In this paper, we present NyLLex, a novel lexical resource derived from books published by Sweden’s largest publisher for easy language texts. NyLLex consists of 6,668 entries, with frequency counts distributed over six reading proficiency levels. We show that NyLLex, with its novel source material aimed at individuals of different reading proficiency levels, can serve as a complement to already existing resources for Swedish.},
  url   = {https://aclanthology.org/2022.lrec-1.141}
}

@InProceedings{uresova-EtAl:2022:LREC,
  author   = {Uresova, Zdenka and Zaczynska, Karolina and Bourgonje, Peter and Fučíková, Eva and Rehm, Georg and Hajic, Jan},
  title   = {Making a Semantic Event-type Ontology Multilingual},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1332--1343},
  abstract = {We present an extension of the SynSemClass Event-type Ontology, originally conceived as a bilingual Czech-English resource. We added German entries to the classes representing the concepts of the ontology. Having a different starting point than the original work (unannotated parallel corpus without links to a valency lexicon and, of course, different existing lexical resources), it was a challenge to adapt the annotation guidelines, the data model and the tools used for the original version. We describe the process and results of working in such a setup. We also show the next steps to adapt the annotation process, data structures and formats and tools necessary to make the addition of a new language in the future more smooth and efficient, and possibly to allow for various teams to work on SynSemClass extensions to many languages concurrently. We also present the latest release which contains the results of adding German, freely available for download as well as for online access.},
  url   = {https://aclanthology.org/2022.lrec-1.142}
}

@InProceedings{kolov-vernerov:2022:LREC,
  author   = {Kolářová, Veronika and Vernerová, Anna},
  title   = {NomVallex: A Valency Lexicon of Czech Nouns and Adjectives},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {1344--1352},
  abstract = {We present NomVallex, a manually annotated valency lexicon of Czech nouns and adjectives. The lexicon is created in the theoretical framework of the Functional Generative Description and based on corpus data. In total, NomVallex 2.0 is comprised of 1027 lexical units contained in 570 lexemes, covering the following part-of-speech and derivational categories: deverbal and deadjectival nouns, and deverbal, denominal, deadjectival and primary adjectives. Valency properties of a lexical unit are captured in a valency frame which is modeled as a sequence of valency slots, supplemented with a list of morphemic forms. In order to make it possible to study the relationship between valency behavior of base words and their derivatives, lexical units of nouns and adjectives in NomVallex are linked to their respective base words, contained either in NomVallex itself or, in case of verbs, in a valency lexicon of Czech verbs called VALLEX. NomVallex enables a comparison of valency properties of a significant number of Czech nominals with their base words, both manually and in an automatic way; as such, we can address the theoretical question of argument inheritance, concentrating on systemic and non-systemic valency behavior.},
  url   = {https://aclanthology.org/2022.lrec-1.143}
}

@InProceedings{aldezabal-arriola-otegi:2022:LREC,
  author   = {Aldezabal, Izaskun and Arriola, Jose Mari and Otegi, Arantxa},
  title   = {TZOS: an Online Terminology Database Aimed at Working on Basque Academic Terminology Collaboratively},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1353--1359},
  abstract = {Terminology databases are highly useful for the dissemination of specialized knowledge. In this paper we present TZOS, an online terminology database to work on Basque academic terminology collaboratively. We show how this resource integrates the Communicative Theory of Terminology, together with the methodological matters, how it is connected with real corpus GARATERM, which terminology issues arise when terms are collected and future perspectives. The main objectives of this work are to develop basic tools to research academic registers and make the terminology collected by expert users available to the community. Even though TZOS has been designed for an educational context, its flexible structure makes possible to extend it also to the professional area. In this way, we have built IZIBI-TZOS which is a Civil Engineering oriented version of TZOS. These resources are already publicly available, and the ongoing work is towards the interlinking with other lexical resources by applying linking data principles.},
  url   = {https://aclanthology.org/2022.lrec-1.144}
}

@InProceedings{klenner-ghring:2022:LREC,
  author   = {Klenner, Manfred and Göhring, Anne},
  title   = {Animacy Denoting German Nouns: Annotation and Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1360--1364},
  abstract = {In this paper, we introduce a gold standard for animacy detection comprising almost 14,500 German nouns that might be used to denote either animate entities or non-animate entities. We present inter-annotator agreement of our crowd-sourced seed annotations (9,000 nouns) and discuss the results of machine learning models applied to this data.},
  url   = {https://aclanthology.org/2022.lrec-1.145}
}

@InProceedings{troiano-EtAl:2022:LREC,
  author   = {Troiano, Enrica and Oberlaender, Laura Ana Maria and Wegge, Maximilian and Klinger, Roman},
  title   = {x-enVENT: A Corpus of Event Descriptions with Experiencer-specific Emotion and Appraisal Annotations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1365--1375},
  abstract = {Emotion classification is often formulated as the task to categorize texts into a predefined set of emotion classes. So far, this task has been the recognition of the emotion of writers and readers, as well as that of entities mentioned in the text. We argue that a classification setup for emotion analysis should be performed in an integrated manner, including the different semantic roles that participate in an emotion episode. Based on appraisal theories in psychology, which treat emotions as reactions to events, we compile an English corpus of written event descriptions. The descriptions depict emotion-eliciting circumstances, and they contain mentions of people who responded emotionally. We annotate all experiencers, including the original author, with the emotions they likely felt. In addition, we link them to the event they found salient (which can be different for different experiencers in a text) by annotating event properties, or appraisals (e.g., the perceived event undesirability, the uncertainty of its outcome). Our analysis reveals patterns in the co-occurrence of people’s emotions in interaction. Hence, this richly-annotated resource provides useful data to study emotions and event evaluations from the perspective of different roles, and it enables the development of experiencer-specific emotion and appraisal classification systems.},

  url   = {https://aclanthology.org/2022.lrec-1.146}
}

@InProceedings{ghring-klenner:2022:LREC,
  author   = {Göhring, Anne and Klenner, Manfred},
  title   = {Polar Quantification of Actor Noun Phrases for German},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1376--1380},
  abstract = {In this paper, we discuss work that strives to measure the degree of negativity - the negative polar load - of noun phrases, especially those denoting actors. Since no gold standard data is available for German for this quantification task, we generated a silver standard and used it to fine-tune a BERT-based intensity regressor. We evaluated the quality of the silver standard empirically and found that our lexicon-based quantification metric showed a strong correlation with human annotators.},
  url   = {https://aclanthology.org/2022.lrec-1.147}
}

@InProceedings{pib-steinberger:2022:LREC,
  author   = {Přibáň, Pavel and Steinberger, Josef},
  title   = {Czech Dataset for Cross-lingual Subjectivity Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1381--1391},
  abstract = {In this paper, we introduce a new Czech subjectivity dataset of 10k manually annotated subjective and objective sentences from movie reviews and descriptions. Our prime motivation is to provide a reliable dataset that can be used with the existing English dataset as a benchmark to test the ability of pre-trained multilingual models to transfer knowledge between Czech and English and vice versa. Two annotators annotated the dataset reaching 0.83 of the Cohen’s K inter-annotator agreement. To the best of our knowledge, this is the first subjectivity dataset for the Czech language. We also created an additional dataset that consists of 200k automatically labeled sentences. Both datasets are freely available for research purposes. Furthermore, we fine-tune five pre-trained BERT-like models to set a monolingual baseline for the new dataset and we achieve 93.56\% of accuracy. We fine-tune models on the existing English dataset for which we obtained results that are on par with the current state-of-the-art results. Finally, we perform zero-shot cross-lingual subjectivity classification between Czech and English to verify the usability of our dataset as the

cross-lingual benchmark. We compare and discuss the cross-lingual and monolingual results and the ability of multilingual models to transfer knowledge between languages.},
  url   = {https://aclanthology.org/2022.lrec-1.148}
}

@InProceedings{ciobotaru-EtAl:2022:LREC,
  author   = {Ciobotaru, Alexandra and Constantinescu, Mihai Vlad and Dinu, Liviu P. and Dumitrescu, Stefan},
  title   = {RED v2: Enhancing RED Dataset for Multi-Label Emotion Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1392--1399},
  abstract = {RED (Romanian Emotion Dataset) is a machine learning-based resource developed for the automatic detection of emotions in Romanian texts, containing single-label annotated tweets with one of the following emotions: joy, fear, sadness, anger and neutral. In this work, we propose REDv2, an open-source extension of RED by adding two more emotions, trust and surprise, and by widening the annotation schema so that the resulted novel dataset is multi-label. We show the overall reliability of our dataset by computing inter-annotator agreements per tweet using a formula suitable for our annotation setup and we aggregate all annotators' opinions into two variants of ground truth, one suitable for multi-label classification and the other suitable for text regression. We propose strong baselines with two transformer models, the Romanian BERT and the multilingual XLM-Roberta model, in both categorical and regression settings.},
  url   = {https://aclanthology.org/2022.lrec-1.149}
}

@InProceedings{xompero-EtAl:2022:LREC,
  author   = {Xompero, Giancarlo and Mastromattei, Michele and Salman, Samir and Giannone, Cristina and Favalli, Andrea and Romagnoli, Raniero and Zanzotto, Fabio Massimo},
  title   = {Every time I fire a conversational designer, the performance of the dialogue system goes down},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {137--145},
  abstract = {Incorporating handwritten domain scripts into neural-based task-oriented dialogue systems may be an effective way to reduce the need for large sets of annotated dialogues. In this paper, we investigate how the use of domain scripts written by conversational designers affects the performance of neural-based

dialogue systems. To support this investigation, we propose the Conversational-Logic-Injection-in-Neural-Network system (CLINN) where domain scripts are coded in semi-logical rules. By using CLINN, we evaluated semi-logical rules produced by a team of differently-skilled conversational designers. We experimented with the Restaurant domain of the MultiWOZ dataset. Results show that external knowledge is extremely important for reducing the need for annotated examples for conversational systems. In fact, rules from conversational designers used in CLINN significantly outperform a state-of-the-art neural-based dialogue system when trained with smaller sets of annotated dialogues.},
  url   = {https://aclanthology.org/2022.lrec-1.15}
}

@InProceedings{ortmann:2022:LREC,
  author   = {Ortmann, Katrin},
  title   = {Fine-Grained Error Analysis and Fair Evaluation of Labeled Spans},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1400--1407},
  abstract = {The traditional evaluation of labeled spans with precision, recall, and F1-score has undesirable effects due to double penalties. Annotations with incorrect label or boundaries count as two errors instead of one, despite being closer to the target annotation than false positives or false negatives. In this paper, new error types are introduced, which more accurately reflect true annotation quality and ensure that every annotation counts only once. An algorithm for error identification in flat and multi-level annotations is presented and complemented with a proposal on how to calculate meaningful precision, recall, and F1-scores based on the more fine-grained error types. The exemplary application to three different annotation tasks (NER, chunking, parsing) shows that the suggested procedure not only prevents double penalties but also allows for a more detailed error analysis, thereby providing more insight into the actual weaknesses of a system.},
  url   = {https://aclanthology.org/2022.lrec-1.150}
}

@InProceedings{epure-hennequin:2022:LREC,
  author   = {Epure, Elena V. and Hennequin, Romain},
  title   = {Probing Pre-trained Auto-regressive Language Models for Named Entity Typing and Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1408--1417},

  abstract = {Multiple works have proposed to probe language models (LMs) for generalization in named entity (NE) typing (NET) and recognition (NER). However, little has been done in this direction for auto-regressive models despite their popularity and potential to express a wide variety of NLP tasks in the same unified format. We propose a new methodology to probe auto-regressive LMs for NET and NER generalization, which draws inspiration from human linguistic behavior, by resorting to meta-learning. We study NEs of various types individually by designing a zero-shot transfer strategy for NET. Then, we probe the model for NER by providing a few examples at inference. We introduce a novel procedure to assess the model's memorization of NEs and report the memorization's impact on the results. Our findings show that: 1) GPT2, a common pre-trained auto-regressive LM, without any fine-tuning for NET or NER, performs the tasksfairly well; 2) name irregularity when common for a NE type could be an effective exploitable cue; 3) the model seems to rely more on NE than contextual cues in few-shot NER; 4) NEs with words absent during LM pre-training are very challenging for both NET and NER.},
  url   = {https://aclanthology.org/2022.lrec-1.151}
}

@InProceedings{vandergoot-mllereberstein-plank:2022:LREC,
  author   = {van der Goot, Rob and Müller-Eberstein, Max and Plank, Barbara},
  title   = {Frustratingly Easy Performance Improvements for Low-resource Setups: A Tale on BERT and Segment Embeddings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1418--1427},
  abstract = {As input representation for each sub-word, the original BERT architecture proposes the sum of the sub-word embedding, position embedding and a segment embedding. Sub-word and position embeddings are well-known and studied, and encode lexical information and word position, respectively. In contrast, segment embeddings are less known and have so far received no attention, despite being ubiquitous in large pre-trained language models. The key idea of segment embeddings is to encode to which of the two sentences (segments) a word belongs to – the intuition is to inform the model about the separation of sentences for the next sentence prediction pre-training task. However, little is known on whether the choice of segment impacts performance. In this work, we try to fill this gap and empirically study the impact of the segment embedding during inference time for a variety of pre-trained embeddings and target tasks. We hypothesize that for single-sentence prediction tasks performance is not affected – neither in mono- nor multilingual setups – while it matters when swapping segment IDs in paired-sentence tasks. To our surprise, this is not the case. Although for classification tasks and monolingual BERT models no large differences are observed, particularly word-level multilingual

prediction tasks are heavily impacted. For low-resource syntactic tasks, we observe impacts of segment embedding and multilingual BERT choice. We find that the default setting for the most used multilingual BERT model underperforms heavily, and a simple swap of the segment embeddings yields an average improvement of 2.5 points absolute LAS score for dependency parsing over 9 different treebanks.},
  url   = {https://aclanthology.org/2022.lrec-1.152}
}

@InProceedings{navarretta-haltruphansen:2022:LREC,
  author   = {Navarretta, Costanza and Haltrup Hansen, Dorte},
  title   = {The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1428--1436},
  abstract = {This paper addresses the semi-automatic annotation of subjects, also called policy areas, in the Danish Parliament Corpus (2009-2017) v.2. Recently, the corpus has been made available through the CLARIN-DK repository, the Danish node of the European CLARIN infrastructure. The paper also contains an analysis of the subjects in the corpus, and a description of multi-label classification experiments act to verify the consistency of the subject annotation and the utility of the corpus for training classifiers on this type of data. The analysis of the corpus comprises an investigation of how often the parliament members addressed each subject and the relation between subjects and gender of the speaker. The classification experiments show that classifiers can determine the two co-occurring subjects of the speeches from the agenda titles with a performance similar to that of human annotators. Moreover, a multilayer perceptron achieved an F1-score of 0.68 on the same task when trained on bag of words vectors obtained from the speeches' lemmas. This is an improvement of more than 0.6 with respect to the baseline, a majority classifier that accounts for the frequency of the classes. The result is promising given the high number of subject combinations (186) and the skewness of the data.},
  url   = {https://aclanthology.org/2022.lrec-1.153}
}

@InProceedings{richardson-wiles:2022:LREC,
  author   = {Richardson, Ashleigh and Wiles, Janet},
  title   = {A Systematic Study Reveals Unexpected Interactions in Pre-Trained Neural Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {1437--1443},
abstract = {A significant challenge in developing translation systems for the world's ~7,000 languages is that very few have sufficient data for state-of-the-art techniques. Transfer learning is a promising direction for low-resource neural machine translation (NMT), but introduces many new variables which are often selected through ablation studies, costly trial-and-error, or niche expertise. When pre-training an NMT system for low-resource translation, the pre-training task is often chosen based on data abundance and similarity to the main task. Factors such as dataset sizes and similarity have typically been analysed independently in previous studies, due to the computational cost associated with systematic studies. However, these factors are not independent. We conducted a three-factor experiment to examine how language similarity, pre-training dataset size and main dataset size interacted in their effect on performance in pre-trained transformer-based low-resource NMT. We replicated the common finding that more data was beneficial in bilingual systems, but also found a statistically significant interaction between the three factors, which reduced the effectiveness of large pre-training datasets for some main task dataset sizes (p-value < 0.0018). The surprising trends identified in these interactions indicate that systematic studies of interactions may be a promising long-term direction for guiding research in low-resource neural methods.},
url = {https://aclanthology.org/2022.lrec-1.154}
}

@InProceedings{ocal-EtAl:2022:LREC1,
author = {Ocal, Mustafa and Perez, Adrian and Radas, Antonela and Finlayson, Mark},
title = {Holistic Evaluation of Automatic TimeML Annotators},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {1444--1453},
abstract = {TimeML is a scheme for representing temporal information (times, events, \& temporal relations) in texts. Although automatic TimeML annotation is challenging, there has been notable progress, with F1s of 0.8--0.9 for events and time detection subtasks, and F1s of 0.5--0.7 for relation extraction. Individually, these subtask results are reasonable, even good, but when combined to generate a full TimeML graph, is overall performance still acceptable? We present a novel suite of eight metrics, combined with a new graph-transformation experimental design, for holistic evaluation of TimeML graphs. We apply these metrics to four automatic TimeML annotation systems (CAEVO, TARSQI, CATENA, and ClearTK). We show that on average 1/3 of the TimeML graphs produced using these systems are inconsistent, and there is on average 1/5 more temporal indeterminacy than the gold-standard. We also show

that the automatically generated graphs are on average 109 edits from the gold-standard, which is 1/3 toward complete replacement. Finally, we show that the relationship individual subtask performance and graph quality is non-linear: small errors in TimeML subtasks result in rapid degradation of final graph quality. These results suggest current automatic TimeML annotators are far from optimal and significant further improvement would be useful.},
  url   = {https://aclanthology.org/2022.lrec-1.155}
}

@InProceedings{gladkoff-EtAl:2022:LREC,
  author   = {Gladkoff, Serge and Sorokina, Irina and Han, Lifeng and Alekseeva, Alexandra},
  title   = {Measuring Uncertainty in Translation Quality Evaluation (TQE)},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1454--1461},
  abstract = {From both human translators (HT) and machine translation (MT) researchers' point of view, translation quality evaluation (TQE) is an essential task. Translation service providers (TSPs) have to deliver large volumes of translations which meet customer specifications with harsh constraints of required quality level in tight time-frames and costs. MT researchers strive to make their models better, which also requires reliable quality evaluation. While automatic machine translation evaluation (MTE) metrics and quality estimation (QE) tools are widely available and easy to access, existing automated tools are not good enough, and human assessment from professional translators (HAP) are often chosen as the golden standard \cite{han-etal-2021-TQA}. Human evaluations, however, are often accused of having low reliability and agreement. Is this caused by subjectivity or statistics is at play? How to avoid the entire text to be checked and be more efficient with TQE from cost and efficiency perspectives, and what is the optimal sample size of the translated text, so as to reliably estimate the translation quality of the entire material? This work carries out such a motivated research to correctly estimate the confidence intervals \cite{Brown_etal2001Interval} depending on the sample size of translated text, e.g. the amount of words or sentences, that needs to be processed on TQE workflow step for confident and reliable evaluation of overall translation quality. The methodology we applied for this work is from Bernoulli Statistical Distribution Modelling (BSDM) and Monte Carlo Sampling Analysis (MCSA).},
  url   = {https://aclanthology.org/2022.lrec-1.156}
}

@InProceedings{altammami-atwell:2022:LREC,
  author   = {Altammami, Shatha and Atwell, Eric},
  title   = {Challenging the Transformer-based models with a

Classical Arabic dataset: Quran and Hadith},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1462--1471},
  abstract = {Transformer-based models showed near-perfect results on several downstream tasks. However, their performance on classical Arabic texts is largely unexplored. To fill this gap, we evaluate monolingual, bilingual, and multilingual state-of-the-art models to detect relatedness between the Quran (Muslim holy book) and the Hadith (Prophet Muhammed teachings), which are complex classical Arabic texts with underlying meanings that require deep human understanding. To do this, we carefully built a dataset of Quran-verse and Hadith-teaching pairs by consulting sources of reputable religious experts. This study presents the methodology of creating the dataset, which we make available on our repository, and discusses the models' performance that calls for the imminent need to explore avenues for improving the quality of these models to capture the semantics in such complex, low-resource texts.},
  url   = {https://aclanthology.org/2022.lrec-1.157}
}

@InProceedings{britton-sarkhel-venugopal:2022:LREC,
  author   = {Britton, William and Sarkhel, Somdeb and Venugopal, Deepak},
  title   = {Question Modifiers in Visual Question Answering},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1472--1479},
  abstract = {Visual Question Answering (VQA) is a challenge problem that can advance AI by integrating several important sub-disciplines including natural language understanding and computer vision. Large VQA datasets that are publicly available for training and evaluation have driven the growth of VQA models that have obtained increasingly larger accuracy scores. However, it is also important to understand how much a model understands the details that are provided in a question. For example, studies in psychology have shown that syntactic complexity places a larger cognitive load on humans. Analogously, we want to understand if models have the perceptual capability to handle modifications to questions. Therefore, we develop a new dataset using Amazon Mechanical Turk where we asked workers to add modifiers to questions based on object properties and spatial relationships. We evaluate this data on LXMERT which is a state-of-the-art model in VQA that focuses more extensively on language processing. Our conclusions indicate that there is a significant negative impact on the performance of the model when the questions are modified to include more detailed

information.},
  url   = {https://aclanthology.org/2022.lrec-1.158}
}

@InProceedings{sosa-sharoff:2022:LREC,
  author   = {Sosa, Jose and Sharoff, Serge},
  title   = {Multimodal Pipeline for Collection of Misinformation Data from Telegram},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1480--1489},
  abstract = {The paper presents the outcomes of AI-COVID19, our project aimed at better understanding of misinformation flow about COVID-19 across social media platforms. The specific focus of the study reported in this paper is on collecting data from Telegram groups which are active in promotion of COVID-related misinformation. Our corpus collected so far contains around 28 million words, from almost one million messages. Given that a substantial portion of misinformation flow in social media is spread via multimodal means, such as images and video, we have also developed a mechanism for utilising such channels via producing automatic transcripts for videos and automatic classification for images into such categories as memes, screenshots of posts and other kinds of images. The accuracy of the image classification pipeline is around 87\%.},
  url   = {https://aclanthology.org/2022.lrec-1.159}
}

@InProceedings{wen-luo-mou:2022:LREC,
  author   = {Wen, Yuqiao and Luo, Guoqing and Mou, Lili},
  title   = {An Empirical Study on the Overlapping Problem of Open-Domain Dialogue Datasets},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {146--153},
  abstract = {Open-domain dialogue systems aim to converse with humans through text, and dialogue research has heavily relied on benchmark datasets. In this work, we observe the overlapping problem in DailyDialog and OpenSubtitles, two popular open-domain dialogue benchmark datasets. Our systematic analysis then shows that such overlapping can be exploited to obtain fake state-of-the-art performance. Finally, we address this issue by cleaning these datasets and setting up a proper data processing procedure for future research.},
  url   = {https://aclanthology.org/2022.lrec-1.16}
}

@InProceedings{xia-EtAl:2022:LREC,
  author   = {Xia, Xinyuan and Xiao, Lu and Yang, Kun and
Wang, Yueyue},
  title   = {Identifying Tension in Holocaust Survivors’
Interview: Code-switching/Code-mixing as Cues},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1490--1495},
  abstract = {In this study, we thrive on finding out how code-
switching and code-mixing (CS/CM) as a linguistic phenomenon could
be a sign of tension in Holocaust survivors' interviews. We first
created an interview corpus (a total of 39 interviews) that contains
manually annotated CS/CM codes (a total of 802 quotations). We then
compared our annotations with the tension places in the corpus. The
tensions are identified by a computational tool. We found that most
of our annotations were captured in the tension places, and it
showed a relatively outstanding performance. The finding implies
that CS/CM can be appropriate cues for detecting tension in this
communication context. Our CS/CM annotated interview corpus is
openly accessible. Aside from annotating and examining CS/CM
occurrences, we annotated silence situations in this open corpus.
Silence is shown to be an indicator of tension in interpersonal
communications. Making this corpus openly accessible, we call for
more research endeavors on tension detection.},
  url   = {https://aclanthology.org/2022.lrec-1.160}
}

@InProceedings{jensen-plank:2022:LREC,
  author   = {Jensen, Kristian Nørgaard and Plank, Barbara},
  title   = {Fine-tuning vs From Scratch: Do Vision \& Language
Models Have Similar Capabilities on Out-of-Distribution Visual
Question Answering?},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1496--1508},
  abstract = {Fine-tuning general-purpose pre-trained models has
become a de-facto standard, also for Vision and Language tasks such
as Visual Question Answering (VQA). In this paper, we take a step
back and ask whether a fine-tuned model has superior linguistic and
reasoning capabilities than a prior state-of-the-art architecture
trained from scratch on the training data alone. We perform a fine-
grained evaluation on out-of-distribution data, including an
analysis on robustness due to linguistic variation (rephrasings).
Our empirical results confirm the benefit of pre-training on overall
performance and rephrasing in particular. But our results also

uncover surprising limitations, particularly for answering questions involving boolean operations. To complement the empirical evaluation, this paper also surveys relevant earlier work on 1) available VQA data sets, 2) models developed for VQA, 3) pre-trained Vision+Language models, and 4) earlier fine-grained evaluation of pre-trained Vision+Language models.},
  url   = {https://aclanthology.org/2022.lrec-1.161}
}

@InProceedings{koeva-stoyanova-kralev:2022:LREC,
  author   = {Koeva, Svetla and Stoyanova, Ivelina and Kralev, Jordan},
  title   = {Multilingual Image Corpus — Towards a Multimodal and Multilingual Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1509--1518},
  abstract = {One of the processing tasks for large multimodal data streams is automatic image description (image classification, object segmentation and classification). Although the number and the diversity of image datasets is constantly expanding, still there is a huge demand for more datasets in terms of variety of domains and object classes covered. The goal of the project Multilingual Image Corpus (MIC 21) is to provide a large image dataset with annotated objects and object descriptions in 24 languages. The Multilingual Image Corpus consists of an Ontology of visual objects (based on WordNet) and a collection of thematically related images whose objects are annotated with segmentation masks and labels describing the ontology classes. The dataset is designed both for image classification and object detection and for semantic segmentation. The main contributions of our work are: a) the provision of large collection of high quality copyright-free images; b) the formulation of the Ontology of visual objects based on WordNet noun hierarchies; c) the precise manual correction of automatic object segmentation within the images and the annotation of object classes; and d) the association of objects and images with extended multilingual descriptions based on WordNet inner- and interlingual relations. The dataset can be used also for multilingual image caption generation, image-to-text alignment and automatic question answering for images and videos.},
  url   = {https://aclanthology.org/2022.lrec-1.162}
}

@InProceedings{kim-EtAl:2022:LREC1,
  author   = {Kim, Jung-Ho and Hwang, Eui Jun and Cho, Sukmin and Lee, Du Hui and Park, Jong},
  title   = {Sign Language Production With Avatar Layering: A Critical Use Case over Rare Words},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1519--1528},
  abstract = {Sign language production (SLP) is the process of generating sign language videos from spoken language expressions. Since sign languages are highly under-resourced, existing vision-based SLP approaches suffer from out-of-vocabulary (OOV) and test-time generalization problems and thus generate low-quality translations. To address these problems, we introduce an avatar-based SLP system composed of a sign language translation (SLT) model and an avatar animation generation module. Our Transformer-based SLT model utilizes two additional strategies to resolve these problems: named entity transformation to reduce OOV tokens and context vector generation using a pretrained language model (e.g., BERT) to reliably train the decoder. Our system is validated on a new Korean-Korean Sign Language (KSL) dataset of weather forecasts and emergency announcements. Our SLT model achieves an 8.77 higher BLEU-4 score and a 4.57 higher ROUGE-L score over those of our baseline model. In a user evaluation, 93.48\% of named entities were successfully identified by participants, demonstrating marked improvement on OOV issues.},
  url   = {https://aclanthology.org/2022.lrec-1.163}
}

@InProceedings{krishnaswamy-EtAl:2022:LREC,
  author   = {Krishnaswamy, Nikhil and Pickard, William and Cates, Brittany and Blanchard, Nathaniel and Pustejovsky, James},
  title   = {The VoxWorld Platform for Multimodal Embodied Agents},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1529--1541},
  abstract = {We present a five-year retrospective on the development of the VoxWorld platform, first introduced as a multimodal platform for modeling motion language, that has evolved into a platform for rapidly building and deploying embodied agents with contextual and situational awareness, capable of interacting with humans in multiple modalities, and exploring their environments. In particular, we discuss the evolution from the theoretical underpinnings of the VoxML modeling language to a platform that accommodates both neural and symbolic inputs to build agents capable of multimodal interaction and hybrid reasoning. We focus on three distinct agent implementations and the functionality needed to accommodate all of them: Diana, a virtual collaborative agent; Kirby, a mobile robot; and BabyBAW, an agent who self-guides its own exploration of the world.},
  url   = {https://aclanthology.org/2022.lrec-1.164}

}

@InProceedings{hossain-sharif-hoque:2022:LREC,
  author   = {Hossain, Eftekhar and Sharif, Omar and Hoque, Mohammed Moshiul},
  title   = {MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1542--1554},
  abstract = {Posting and sharing memes have become a powerful expedient of expressing opinions on social media in recent days. Analysis of sentiment from memes has gained much attention to researchers due to its substantial implications in various domains like finance and politics. Past studies on sentiment analysis of memes have primarily been conducted in English, where low-resource languages gain little or no attention. However, due to the proliferation of social media usage in recent years, sentiment analysis of memes is also a crucial research issue in low resource languages. The scarcity of benchmark datasets is a significant barrier to performing multimodal sentiment analysis research in resource-constrained languages like Bengali. This paper presents a novel multimodal dataset (named MemoSen) for Bengali containing 4417 memes with three annotated labels positive, negative, and neutral. A detailed annotation guideline is provided to facilitate further resource development in this domain. Additionally, a set of experiments are carried out on MemoSen by constructing twelve unimodal (i.e., visual, textual) and ten multimodal (image+text) models. The evaluation exhibits that the integration of multimodal information significantly improves (about 1.2\%) the meme sentiment classification compared to the unimodal counterparts and thus elucidate the novel aspects of multimodality.},
  url   = {https://aclanthology.org/2022.lrec-1.165}
}

@InProceedings{ivanko-EtAl:2022:LREC,
  author   = {Ivanko, Denis and Axyonov, Alexandr and Ryumin, Dmitry and Kashevnik, Alexey and Karpov, Alexey},
  title   = {RUSAVIC Corpus: Russian Audio-Visual Speech in Cars},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1555--1559},
  abstract = {We present a new audio-visual speech corpus (RUSAVIC) recorded in a car environment and designed for noise-robust speech recognition. Our goal was to produce a speech corpus which is natural (recorded in real driving conditions), controlled (providing

different SNR levels by windows open/closed, moving/parked vehicle, etc.), and adequate size (the amount of data is enough to train state-of-the-art NN approaches). We focus on the problem of audio-visual speech recognition: with the use of automated lip-reading to improve the performance of audio-based speech recognition in the presence of severe acoustic noise caused by road traffic. We also describe the equipment and procedures used to create RUSAVIC corpus. Data are collected in a synchronous way through several smartphones located at different angles and equipped with FullHD video camera and microphone. The corpus includes the recordings of 20 drivers with minimum of 10 recording sessions for each. Besides providing a detailed description of the dataset and its collection pipeline, we evaluate several popular audio and visual speech recognition methods and present a set of baseline recognition results. At the moment RUSAVIC is a unique audio-visual corpus for the Russian language that is recorded in-the-wild condition and we make it publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.166}
}

@InProceedings{challant-filhol:2022:LREC,
  author   = {Challant, Camille and Filhol, Michael},
  title   = {A First Corpus of AZee Discourse Expressions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1560--1565},
  abstract = {This paper presents a corpus of AZee discourse expressions, i.e. expressions which formally describe Sign Language utterances of any length using the AZee approach and language. The construction of this corpus had two main goals: a first reference corpus for AZee, and a test of its coverage on a significant sample of real-life utterances. We worked on productions from an existing corpus, namely the "40 breves", containing an hour of French Sign Language. We wrote the corresponding AZee discourse expressions for the entire video content, i.e. expressions capturing the forms produced by the signers and their associated meaning by combining known production rules, a basic building block for these expressions. These are made available as a version 2 extension of the "40 breves". We explain the way in which these expressions can be built, present the resulting corpus and set of production rules used, and perform first measurements on it. We also propose an evaluation of our corpus: for one hour of discourse, AZee allows to describe 94\% of it, while ongoing studies are increasing this coverage. This corpus offers a lot of future prospects, for instance concerning synthesis with virtual signers, machine translation or formal grammars for Sign Language.},
  url   = {https://aclanthology.org/2022.lrec-1.167}
}

@InProceedings{lebron-EtAl:2022:LREC,

  author   = {Lebron, Luis and Graham, Yvette and McGuinness, Kevin and Kouramas, Konstantinos and O'Connor, Noel E.},
  title   = {BERTHA: Video Captioning Evaluation Via Transfer-Learned Human Assessment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1566--1575},
  abstract = {Evaluating video captioning systems is a challenging task as there are multiple factors to consider; for instance: the fluency of the caption, multiple actions happening in a single scene, and the human bias of what is considered important. Most metrics try to measure how similar the system generated captions are to a single or a set of human-annotated captions. This paper presents a new method based on a deep learning model to evaluate these systems. The model is based on BERT, which is a language model that has been shown to work well in multiple NLP tasks. The aim is for the model to learn to perform an evaluation similar to that of a human. To do so, we use a dataset that contains human evaluations of system generated captions. The dataset consists of the human judgments of the captions produces by the system participating in various years of the TRECVid video to text task. BERTHA obtain favourable results, outperforming the commonly used metrics in some setups.},
  url   = {https://aclanthology.org/2022.lrec-1.168}
}

@InProceedings{brutti-EtAl:2022:LREC,
  author   = {Brutti, Richard and Donatelli, Lucia and Lai, Kenneth and Pustejovsky, James},
  title   = {Abstract Meaning Representation for Gesture},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1576--1583},
  abstract = {This paper presents Gesture AMR, an extension to Abstract Meaning Representation (AMR), that captures the meaning of gesture. In developing Gesture AMR, we consider how gesture form and meaning relate; how gesture packages meaning both independently and in interaction with speech; and how the meaning of gesture is temporally and contextually determined. Our case study for developing Gesture AMR is a focused human-human shared task to build block structures. We develop an initial taxonomy of gesture act relations that adheres to AMR’s existing focus on predicate-argument structure while integrating meaningful elements unique to gesture. Pilot annotation shows Gesture AMR to be more challenging than standard AMR, and illustrates the need for more work on representation of dialogue and multimodal meaning. We discuss

challenges of adapting an existing meaning representation to non-speech-based modalities and outline several avenues for expanding Gesture AMR.},
  url   = {https://aclanthology.org/2022.lrec-1.169}
}

@InProceedings{gamba-EtAl:2022:LREC,
  author   = {Gamba, Federica and Frontini, Francesca and Broeder, Daan and Monachini, Monica},
  title   = {Language Technologies for the Creation of Multilingual Terminologies. Lessons Learned from the SSHOC Project},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {154--163},
  abstract = {This paper is framed in the context of the SSHOC project and aims at exploring how Language Technologies can help in promoting and facilitating multilingualism in the Social Sciences and Humanities (SSH). Although most SSH researchers produce culturally and societally relevant work in their local languages, metadata and vocabularies used in the SSH domain to describe and index research data are currently mostly in English. We thus investigate Natural Language Processing and Machine Translation approaches in view of providing resources and tools to foster multilingual access and discovery to SSH content across different languages. As case studies, we create and deliver as freely, openly available data a set of multilingual metadata concepts and an automatically extracted multilingual Data Stewardship terminology. The two case studies allow as well to evaluate performances of state-of-the-art tools and to derive a set of recommendations as to how best apply them. Although not adapted to the specific domain, the employed tools prove to be a valid asset to translation tasks. Nonetheless, validation of results by domain experts proficient in the language is an unavoidable phase of the whole workflow.},
  url   = {https://aclanthology.org/2022.lrec-1.17}
}

@InProceedings{kuzman-rupnik-ljubei:2022:LREC,
  author   = {Kuzman, Taja and Rupnik, Peter and Ljubešić, Nikola},
  title   = {The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1584--1594},
  abstract = {This paper presents a new training dataset for automatic genre identification GINCO, which is based on 1,125

crawled Slovenian web documents that consist of 650,000 words. Each document was manually annotated for genre with a new annotation schema that builds upon existing schemata, having primarily clarity of labels and inter-annotator agreement in mind. The dataset consists of various challenges related to web-based data, such as machine translated content, encoding errors, multiple contents presented in one document etc., enabling evaluation of classifiers in realistic conditions. The initial machine learning experiments on the dataset show that (1) pre-Transformer models are drastically less able to model the phenomena, with macro F1 metrics ranging around 0.22, while Transformer-based models achieve scores of around 0.58, and (2) multilingual Transformer models work as well on the task as the monolingual models that were previously proven to be superior to multilingual models on standard NLP tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.170}
}

@InProceedings{laperrire-EtAl:2022:LREC,
  author   = {Laperrière, Gaëlle and Pelloin, Valentin and Caubrière, Antoine and mdhaffar, salima and Camelin, Nathalie and Ghannay, Sahar and Jabaian, Bassam and Estève, Yannick},
  title   = {The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning: data updates, training and evaluation tools},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1595--1602},
  abstract = {With the emergence of neural end-to-end approaches for spoken language understanding (SLU), a growing number of studies have been presented during these last three years on this topic. The major part of these works addresses the spoken language understanding domain through a simple task like speech intent detection. In this context, new benchmark datasets have also been produced and shared with the community related to this task. In this paper, we focus on the French MEDIA SLU dataset, distributed since 2005 and used as a benchmark dataset for a large number of research works. This dataset has been shown as being the most challenging one among those accessible to the research community. Distributed by ELRA, this corpus is free for academic research since 2019. Unfortunately, the MEDIA dataset is not really used beyond the French research community. To facilitate its use, a complete recipe, including data preparation, training and evaluation scripts, has been built and integrated to SpeechBrain, an already popular open-source and all-in-one conversational AI toolkit based on PyTorch. This recipe is presented in this paper. In addition, based on the feedback of some researchers who have worked on this dataset for several years, some corrections have been brought to the initial manual annotation: the new version of the data will also be integrated into the ELRA catalogue, as the original one. More, a significant amount of data collected during the construction of the

MEDIA corpus in the 2000s was never used until now: we present the first results reached on this subset – also included in the MEDIA SpeechBrain recipe – , that will be used for now as the MEDIA test2. Last, we discuss evaluation issues.},
  url   = {https://aclanthology.org/2022.lrec-1.171}
}

@InProceedings{urbizu-EtAl:2022:LREC,
  author   = {Urbizu, Gorka and San Vicente, Iñaki and Saralegi, Xabier and Agerri, Rodrigo and Soroa, Aitor},
  title   = {BasqueGLUE: A Natural Language Understanding Benchmark for Basque},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1603--1612},
  abstract = {Natural Language Understanding (NLU) technology has improved significantly over the last few years and multitask benchmarks such as GLUE are key to evaluate this improvement in a robust and general way. These benchmarks take into account a wide and diverse set of NLU tasks that require some form of language understanding, beyond the detection of superficial, textual clues. However, they are costly to develop and language-dependent, and therefore they are only available for a small number of languages. In this paper, we present BasqueGLUE, the first NLU benchmark for Basque, a less-resourced language, which has been elaborated from previously existing datasets and following similar criteria to those used for the construction of GLUE and SuperGLUE. We also report the evaluation of two state-of-the-art language models for Basque on BasqueGLUE, thus providing a strong baseline to compare upon. BasqueGLUE is freely available under an open license.},
  url   = {https://aclanthology.org/2022.lrec-1.172}
}

@InProceedings{stefanovitch-piskorski-kharazi:2022:LREC,
  author   = {Stefanovitch, Nicolas and Piskorski, Jakub and Kharazi, Sopho},
  title   = {Resources and Experiments on Sentiment Classification for Georgian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1613--1621},
  abstract = {This paper presents, to the best of our knowledge, the first ever publicly available annotated dataset for sentiment classification and semantic polarity dictionary for Georgian. The characteristics of these resources and the process of their creation are described in detail. The results of various experiments on the

performance of both lexicon- and machine learning-based models for Georgian sentiment classification are also reported. Both 3-label (positive, neutral, negative) and 4-label settings (same labels + mixed) are considered. The machine learning models explored include, i.a., logistic regression, SVMs, and transformed-based models. We also explore transfer learning- and translation-based (to a well-supported language) approaches. The obtained results for Georgian are on par with the state-of-the-art results in sentiment classification for well studied languages when using training data of comparable size.},
  url   = {https://aclanthology.org/2022.lrec-1.173}
}

@InProceedings{zmandar-EtAl:2022:LREC,
  author   = {ZMANDAR, Nadhem and Daudert, Tobias and Ahmadi, Sina and El-Haj, Mahmoud and Rayson, Paul},
  title   = {CoFiF Plus: A French Financial Narrative Summarisation Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1622--1639},
  abstract = {Natural Language Processing is increasingly being applied in the finance and business industry to analyse the text of many different types of financial documents. Given the increasing growth of firms around the world, the volume of financial disclosures and financial texts in different languages and forms is increasing sharply and therefore the study of language technology methods that automatically summarise content has grown rapidly into a major research area. Corpora for financial narrative summarisation exists in English, but there is a significant lack of financial text resources in the French language. To remedy this, we present CoFiF Plus, the first French financial narrative summarisation dataset providing a comprehensive set of financial text written in French. The dataset has been extracted from French financial reports published in PDF file format. It is composed of 1,703 reports from the most capitalised companies in France (Euronext Paris) covering a time frame from 1995 to 2021. This paper describes the collection, annotation and validation of the financial reports and their summaries. It also describes the dataset and gives the results of some baseline summarisers. Our datasets will be openly available upon the acceptance of the paper.},
  url   = {https://aclanthology.org/2022.lrec-1.174}
}

@InProceedings{calizzano-EtAl:2022:LREC,
  author   = {Calizzano, Rémi and Ostendorff, Malte and Ruan, Qian and Rehm, Georg},
  title   = {Generating Extended and Multilingual Summaries with Pre-trained Transformers},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1640--1650},
  abstract = {Almost all summarisation methods and datasets focus on a single language and short summaries. We introduce a new dataset called WikinewsSum for English, German, French, Spanish, Portuguese, Polish, and Italian summarisation tailored for extended summaries of approx. 11 sentences. The dataset comprises 39,626 summaries which are news articles from Wikinews and their sources. We compare three multilingual transformer models on the extractive summarisation task and three training scenarios on which we fine-tune mT5 to perform abstractive summarisation. This results in strong baselines for both extractive and abstractive summarisation on WikinewsSum. We also show how the combination of an extractive model with an abstractive one can be used to create extended abstractive summaries from long input documents. Finally, our results show that fine-tuning mT5 on all the languages combined significantly improves the summarisation performance on low-resource languages.},
  url   = {https://aclanthology.org/2022.lrec-1.175}
}

@InProceedings{martin-EtAl:2022:LREC,
  author   = {Martin, Louis and Fan, Angela and de la Clergerie, Éric and Bordes, Antoine and Sagot, Benoît},
  title   = {MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1651--1664},
  abstract = {Progress in sentence simplification has been hindered by a lack of labeled parallel simplification data, particularly in languages other than English. We introduce MUSS, a Multilingual Unsupervised Sentence Simplification system that does not require labeled simplification data. MUSS uses a novel approach to sentence simplification that trains strong models using sentence-level paraphrase data instead of proper simplification data. These models leverage unsupervised pretraining and controllable generation mechanisms to flexibly adjust attributes such as length and lexical complexity at inference time. We further present a method to mine such paraphrase data in any language from Common Crawl using semantic sentence embeddings, thus removing the need for labeled data. We evaluate our approach on English, French, and Spanish simplification benchmarks and closely match or outperform the previous best supervised results, despite not using any labeled simplification data. We push the state of the art further by incorporating labeled simplification data.},
  url   = {https://aclanthology.org/2022.lrec-1.176}

}

@InProceedings{honnavalli-EtAl:2022:LREC,
  author   = {Honnavalli, Samhita and Parekh, Aesha and Ou, Lily and Groenwold, Sophie and Levy, Sharon and Ordonez, Vicente and Wang, William Yang},
  title   = {Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1665--1670},
  abstract = {Women are often perceived as junior to their male counterparts, even within the same job titles. While there has been significant progress in the evaluation of gender bias in natural language processing (NLP), existing studies seldom investigate how biases toward gender groups change when compounded with other societal biases. In this work, we investigate how seniority impacts the degree of gender bias exhibited in pretrained neural generation models by introducing a novel framework for probing compound bias. We contribute a benchmark robustness-testing dataset spanning two domains, U.S. senatorship and professorship, created using a distant-supervision method. Our dataset includes human-written text with underlying ground truth and paired counterfactuals. We then examine GPT-2 perplexity and the frequency of gendered language in generated text. Our results show that GPT-2 amplifies bias by considering women as junior and men as senior more often than the ground truth in both domains. These results suggest that NLP applications built using GPT-2 may harm women in professional capacities.},
  url   = {https://aclanthology.org/2022.lrec-1.177}
}

@InProceedings{tamburini:2022:LREC,
  author   = {Tamburini, Fabio},
  title   = {Combining ELECTRA and Adaptive Graph Encoding for Frame Identification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1671--1679},
  abstract = {This paper presents contributions in two directions: first we propose a new system for Frame Identification (FI), based on pre-trained text encoders trained discriminatively and graphs embedding, producing state of the art performance and, second, we take in consideration all the extremely different procedures used to evaluate systems for this task performing a complete evaluation over two benchmarks and all possible splits and cleaning procedures used

in the FI literature.},
  url   = {https://aclanthology.org/2022.lrec-1.178}
}

@InProceedings{garsoler-labeau-clavel:2022:LREC,
  author   = {Garí Soler, Aina and Labeau, Matthieu and Clavel, Chloé},
  title   = {Polysemy in Spoken Conversations and Written Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1680--1690},
  abstract = {Our discourses are full of potential lexical ambiguities, due in part to the pervasive use of words having multiple senses. Sometimes, one word may even be used in more than one sense throughout a text. But, to what extent is this true for different kinds of texts? Does the use of polysemous words change when a discourse involves two people, or when speakers have time to plan what to say? We investigate these questions by comparing the polysemy level of texts of different nature, with a focus on spontaneous spoken dialogs; unlike previous work which examines solely scripted, written, monolog-like data. We compare multiple metrics that presuppose different conceptualizations of text polysemy, i.e., they consider the observed or the potential number of senses of words, or their sense distribution in a discourse. We show that the polysemy level of texts varies greatly depending on the kind of text considered, with dialog and spoken discourses having generally a higher polysemy level than written monologs. Additionally, our results emphasize the need for relaxing the popular "one sense per discourse" hypothesis.},
  url   = {https://aclanthology.org/2022.lrec-1.179}
}

@InProceedings{schulder-hanke:2022:LREC,
  author   = {Schulder, Marc and Hanke, Thomas},
  title   = {How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {164--173},
  abstract = {The publication of resources for minority languages requires a balance between making data open and accessible and respecting the rights and needs of its language community. The FAIR principles were introduced as a guide to good open data practices and they have since been complemented by the CARE principles for indigenous data governance. This article describes how the DGS Corpus implemented these principles and how the two sets of

principles affected each other. The DGS Corpus is a large collection of recordings of members of the deaf community in Germany communicating in their primary language, German Sign Language (DGS); it was created to be both as a resource for linguistic research and as a record of the life experiences of deaf people in Germany. The corpus was designed with CARE in mind to respect and empower the language community and FAIR data publishing was used to enhance its usefulness as a scientific resource.},
  url   = {https://aclanthology.org/2022.lrec-1.18}
}

@InProceedings{batanovi-milievipetrovi:2022:LREC,
  author   = {Batanović, Vuk and Miličević Petrović, Maja},
  title   = {Cross-Level Semantic Similarity for Serbian Newswire Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1691--1699},
  abstract = {Cross-Level Semantic Similarity (CLSS) is a measure of the level of semantic overlap between texts of different lengths. Although this problem was formulated almost a decade ago, research on it has been sparse, and limited exclusively to the English language. In this paper, we present the first CLSS dataset in another language, in the form of CLSS.news.sr – a corpus of 1000 phrase-sentence and 1000 sentence-paragraph newswire text pairs in Serbian, manually annotated with fine-grained semantic similarity scores using a 0–4 similarity scale. We describe the methodology of data collection and annotation, and compare the resulting corpus to its preexisting counterpart in English, SemEval CLSS, following up with a preliminary linguistic analysis of the newly created dataset. State-of-the-art pre-trained language models are then fine-tuned and evaluated on the CLSS task in Serbian using the produced data, and their settings and results are discussed. The CLSS.news.sr corpus and the guidelines used in its creation are made publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.180}
}

@InProceedings{jindal-EtAl:2022:LREC,
  author   = {Jindal, Ishan and Rademaker, Alexandre and Ulewicz, Michał and Linh, Ha and Nguyen, Huyen and Tran, Khoi-Nguyen and Zhu, Huaiyu and Li, Yunyao},
  title   = {Universal Proposition Bank 2.0},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1700--1711},

  abstract = {Semantic role labeling (SRL) represents the meaning of a sentence in the form of predicate-argument structures. Such shallow semantic analysis is helpful in a wide range of downstream NLP tasks and real-world applications. As treebanks enabled the development of powerful syntactic parsers, the accurate predicate-argument analysis demands training data in the form of propbanks. Unfortunately, most languages simply do not have corresponding propbanks due to the high cost required to construct such resources. To overcome such challenges, Universal Proposition Bank 1.0 (UP1.0) was released in 2017, with high-quality propbank data generated via a two-stage method exploiting monolingual SRL and multilingual parallel data. In this paper, we introduce Universal Proposition Bank 2.0 (UP2.0), with significant enhancements over UP1.0: (1) propbanks with higher quality by using a state-of-the-art monolingual SRL and improved auto-generation of annotations; (2) expanded language coverage (from 7 to 9 languages); (3) span annotation for the decoupling of syntactic analysis; and (4) Gold data for a subset of the languages. We also share our experimental results that confirm the significant quality improvements of the generated propbanks. In addition, we present a comprehensive experimental evaluation on how different implementation choices impact the quality of the resulting data. We release these resources to the research community and hope to encourage more research on cross-lingual SRL.},
  url   = {https://aclanthology.org/2022.lrec-1.181}
}

@InProceedings{hollenstein-barrett-bjrnsdttir:2022:LREC,
  author   = {Hollenstein, Nora and Barrett, Maria and Björnsdóttir, Marina},
  title   = {The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1712--1720},
  abstract = {Eye movement recordings from reading are one of the richest signals of human language processing. Corpora of eye movements during reading of contextualized running text is a way of making such records available for natural language processing purposes. Such corpora already exist in some languages. We present CopCo, the Copenhagen Corpus of eye tracking recordings from natural reading of Danish texts. It is the first eye tracking corpus of its kind for the Danish language. CopCo includes 1,832 sentences with 34,897 tokens of Danish text extracted from a collection of speech manuscripts. This first release of the corpus contains eye tracking data from 22 participants. It will be extended continuously with more participants and texts from other genres. We assess the data quality of the recorded eye movements and find that the extracted features are in line with related research. The dataset available here: https://osf.io/ud8s5/.},

  url   = {https://aclanthology.org/2022.lrec-1.182}
}

@InProceedings{weise-mcneill-levitan:2022:LREC,
  author   = {Weise, Andreas and McNeill, Matthew and Levitan, Rivka},
  title   = {The Brooklyn Multi-Interaction Corpus for Analyzing Variation in Entrainment Behavior},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1721--1731},
  abstract = {We present the Brooklyn Multi-Interaction Corpus (B-MIC), a collection of dyadic conversations designed to identify speaker traits and conversation contexts that cause variations in entrainment behavior. B-MIC pairs each participant with multiple partners for an object placement game and open-ended discussions, as well as with a Wizard of Oz for a baseline of their speech. In addition to fully transcribed recordings, it includes demographic information and four completed psychological questionnaires for each subject and turn annotations for perceived emotion and acoustic outliers. This enables the study of speakers' entrainment behavior in different contexts and the sources of variation in this behavior. In this paper, we introduce B-MIC and describe our collection, annotation, and preprocessing methodologies. We report a preliminary study demonstrating varied entrainment behavior across different conversation types and discuss the rich potential for future work on the corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.183}
}

@InProceedings{miletic-EtAl:2022:LREC,
  author   = {Miletic, Aleksandra and Benzitoun, Christophe and Cislaru, Georgeta and Herrera-Yanez, Santiago},
  title   = {Pro-TEXT: an Annotated Corpus of Keystroke Logs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1732--1739},
  abstract = {Pro-TEXT is a corpus of keystroke logs written in French. Keystroke logs are recordings of the writing process executed through a keyboard, which keep track of all actions taken by the writer (character additions, deletions, substitutions). As such, the Pro-TEXT corpus offers new insights into text genesis and underlying cognitive processes from the production perspective. A subset of the corpus is linguistically annotated with parts of speech, lemmas and syntactic dependencies, making it suitable for the study of interactions between linguistic and behavioural aspects

of the writing process. The full corpus contains 202K tokens, while the annotated portion is currently 30K tokens large. The annotated content is progressively being made available in a database-like CSV format and in CoNLL format, and the work on an HTML-based visualisation tool is currently under way. To the best of our knowledge, Pro-TEXT is the first corpus of its kind in French.},
  url   = {https://aclanthology.org/2022.lrec-1.184}
}

@InProceedings{bonetti-EtAl:2022:LREC,
  author   = {Bonetti, Federico and Leonardelli, Elisa and Trotta, Daniela and Guarasci, Raffaele and Tonelli, Sara},
  title   = {Work Hard, Play Hard: Collecting Acceptability Annotations through a 3D Game},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1740--1750},
  abstract = {Corpus-based studies on acceptability judgements have always stimulated the interest of researchers, both in theoretical and computational fields. Some approaches focused on spontaneous judgements collected through different types of tasks, others on data annotated through crowd-sourcing platforms, still others relied on expert annotated data available from the literature. The release of CoLA corpus, a large-scale corpus of sentences extracted from linguistic handbooks as examples of acceptable/non acceptable phenomena in English, has revived interest in the reliability of judgements of linguistic experts vs. non-experts. Several issues are still open. In this work, we contribute to this debate by presenting a 3D video game that was used to collect acceptability judgments on Italian sentences. We analyse the resulting annotations in terms of agreement among players and by comparing them with experts' acceptability judgments. We also discuss different game settings to assess their impact on participants' motivation and engagement. The final dataset containing 1,062 sentences, which were selected based on majority voting, is released for future research and comparisons.},
  url   = {https://aclanthology.org/2022.lrec-1.185}
}

@InProceedings{lapshinovakoltunski-popovi-koponen:2022:LREC,
  author   = {Lapshinova-Koltunski, Ekaterina and Popović, Maja and Koponen, Maarit},
  title   = {DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {1751--1760},
  abstract = {This paper describes a new corpus of human translations which contains both professional and students translations. The data consists of English sources -- texts from news and reviews -- and their translations into Russian and Croatian, as well as of the subcorpus containing translations of the review texts into Finnish. All target languages represent mid-resourced and less or mid-investigated ones. The corpus will be valuable for studying variation in translation as it allows a direct comparison between human translations of the same source texts. The corpus will also be a valuable resource for evaluating machine translation systems. We believe that this resource will facilitate understanding and improvement of the quality issues in both human and machine translation. In the paper, we describe how the data was collected, provide information on translator groups and summarise the differences between the human translations at hand based on our preliminary results with shallow features.},
  url   = {https://aclanthology.org/2022.lrec-1.186}
}

@InProceedings{jahan-EtAl:2022:LREC,
  author   = {Jahan, Md Saroar and Beddiar, Djamila Romaissa and Oussalah, Mourad and Mohamed, Muhidin},
  title   = {Data Expansion Using WordNet-based Semantic Expansion and Word Disambiguation for Cyberbullying Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1761--1770},
  abstract = {Automatic identification of cyberbullying from textual content is known to be a challenging task. The challenges arise from the inherent structure of cyberbullying and the lack of labeled large-scale corpus, enabling efficient machine-learning-based tools including neural networks. This paper advocates a data augmentation-based approach that could enhance the automatic detection of cyberbullying in social media texts. We use both word sense disambiguation and synonymy relation in WordNet lexical database to generate coherent equivalent utterances of cyberbullying input data. The disambiguation and semantic expansion are intended to overcome the inherent limitations of social media posts, such as an abundance of unstructured constructs and limited semantic content. Besides, to test the feasibility, a novel protocol has been employed to collect cyberbullying traces data from AskFm forum, where about a 10K-size dataset has been manually labeled. Next, the problem of cyberbullying identification is viewed as a binary classification problem using an elaborated data augmentation strategy and an appropriate classifier. For the latter, a Convolutional Neural Network (CNN) architecture with FastText and BERT was put forward, whose results were compared against commonly employed Naïve Bayes (NB) and Logistic Regression (LR) classifiers with and without data augmentation. The research outcomes were

promising and yielded almost 98.4\% of classifier accuracy, an improvement of more than 4\% over baseline results.},
  url   = {https://aclanthology.org/2022.lrec-1.187}
}

@InProceedings{polk-EtAl:2022:LREC,
  author   = {Polák, Peter and Singh, Muskaan and Nedoluzhko, Anna and Bojar, Ondřej},
  title   = {ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1771--1779},
  abstract = {Summarization is a challenging problem, and even more challenging is to manually create, correct, and evaluate the summaries. The severity of the problem grows when the inputs are multi-party dialogues in a meeting setup. To facilitate the research in this area, we present ALIGNMEET, a comprehensive tool for meeting annotation, alignment, and evaluation. The tool aims to provide an efficient and clear interface for fast annotation while mitigating the risk of introducing errors. Moreover, we add an evaluation mode that enables a comprehensive quality evaluation of meeting minutes. To the best of our knowledge, there is no such tool available. We release the tool as open source. It is also directly installable from PyPI.},
  url   = {https://aclanthology.org/2022.lrec-1.188}
}

@InProceedings{bayerl-EtAl:2022:LREC,
  author   = {Bayerl, Sebastian and Wolff von Gudenberg, Alexander and Hönig, Florian and Noeth, Elmar and Riedhammer, Korbinian},
  title   = {KSoF: The Kassel State of Fluency Dataset – A Therapy Centered Dataset of Stuttering},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1780--1787},
  abstract = {Stuttering is a complex speech disorder that negatively affects an individual’s ability to communicate effectively. Persons who stutter (PWS) often suffer considerably under the condition and seek help through therapy. Fluency shaping is a therapy approach where PWSs learn to modify their speech to help them to overcome their stutter. Mastering such speech techniques takes time and practice, even after therapy. Shortly after therapy, success is evaluated highly, but relapse rates are high. To be able to monitor speech behavior over a long time, the

ability to detect stuttering events and modifications in speech could help PWSs and speech pathologists to track the level of fluency. Monitoring could create the ability to intervene early by detecting lapses in fluency. To the best of our knowledge, no public dataset is available that contains speech from people who underwent stuttering therapy that changed the style of speaking. This work introduces the Kassel State of Fluency (KSoF), a therapy-based dataset containing over 5500 clips of PWSs. The clips were labeled with six stuttering-related event types: blocks, prolongations, sound repetitions, word repetitions, interjections, and – specific to therapy – speech modifications. The audio was recorded during therapy sessions at the Institut der Kasseler Stottertherapie. The data will be made available for research purposes upon request.},
  url   = {https://aclanthology.org/2022.lrec-1.189}
}

@InProceedings{basile-EtAl:2022:LREC,
  author   = {Basile, Valerio and Bosco, Cristina and Fell, Michael and Patti, Viviana and Varvara, Rossella},
  title   = {Italian NLP for Everyone: Resources and Models from EVALITA to the European Language Grid},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {174--180},
  abstract = {The European Language Grid enables researchers and practitioners to easily distribute and use NLP resources and models, such as corpora and classifiers. We describe in this paper how, during the course of our EVALITA4ELG project, we have integrated datasets and systems for the Italian language. We show how easy it is to use the integrated systems, and demonstrate in case studies how seamless the application of the platform is, providing Italian NLP for everyone.},
  url   = {https://aclanthology.org/2022.lrec-1.19}
}

@InProceedings{guibon-EtAl:2022:LREC,
  author   = {Guibon, Gaël and Lefeuvre, Luce and Labeau, Matthieu and Clavel, Chloé},
  title   = {EZCAT: an Easy Conversation Annotation Tool},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1788--1797},
  abstract = {Users generate content constantly, leading to new data requiring annotation. Among this data, textual conversations are created every day and come with some specificities: they are mostly private through instant messaging applications, requiring the

conversational context to be labeled. These specificities led to several annotation tools dedicated to conversation, and mostly dedicated to dialogue tasks, requiring complex annotation schemata, not always customizable and not taking into account conversation-level labels. In this paper, we present EZCAT, an easy-to-use interface to annotate conversations in a two-level configurable schema, leveraging message-level labels and conversation-level labels. Our interface is characterized by the voluntary absence of a server and accounts management, enhancing its availability to anyone, and the control over data, which is crucial to confidential conversations. We also present our first usage of EZCAT along with our annotation schema we used to annotate confidential customer service conversations. EZCAT is freely available at https://gguibon.github.io/ezcat.},
  url   = {https://aclanthology.org/2022.lrec-1.190}
}

@InProceedings{dobrovoljc:2022:LREC,
  author   = {Dobrovoljc, Kaja},
  title   = {Spoken Language Treebanks in Universal Dependencies: an Overview},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1798--1806},
  abstract = {Given the benefits of syntactically annotated collections of transcribed speech in spoken language research and applications, many spoken language treebanks have been developed in the last decades, with divergent annotation schemes posing important limitations to cross-resource explorations, such as comparing data across languages, grammatical frameworks, and language domains. As a consequence, there has been a growing number of spoken language treebanks adopting the Universal Dependencies (UD) annotation scheme, aimed at cross-linguistically consistent morphosyntactic annotation. In view of the non-central role of spoken language data within the scheme and with little in-domain consolidation to date, this paper presents a comparative overview of spoken language treebanks in UD to support cross-treebank data explorations on the one hand, and encourage further treebank harmonization on the other. Our results show that the spoken language treebanks differ considerably with respect to the inventory and the format of transcribed phenomena, as well as the principles adopted in their morphosyntactic annotation. This is particularly true for the dependency annotation of speech disfluencies, where conflicting data annotations suggest an underspecification of the guidelines pertaining to speech repairs in general and the reparandum dependency relation in particular.},
  url   = {https://aclanthology.org/2022.lrec-1.191}
}

@InProceedings{vanroy-macken:2022:LREC,

  author   = {Vanroy, Bram and Macken, Lieve},
  title   = {LeConTra: A Learner Corpus of English-to-Dutch News Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1807--1816},
  abstract = {We present LeConTra, a learner corpus consisting of English-to-Dutch news translations enriched with translation process data. Three students of a Master's programme in Translation were asked to translate 50 different English journalistic texts of approximately 250 tokens each. Because we also collected translation process data in the form of keystroke logging, our dataset can be used as part of different research strands such as translation process research, learner corpus research, and corpus-based translation studies. Reference translations, without process data, are also included. The data has been manually segmented and tokenized, and manually aligned at both segment and word level, leading to a high-quality corpus with token-level process data. The data is freely accessible via the Translation Process Research DataBase, which emphasises our commitment of distributing our dataset. The tool that was built for manual sentence segmentation and tokenization, Mantis, is also available as an open-source aid for data processing.},
  url   = {https://aclanthology.org/2022.lrec-1.192}
}

@InProceedings{hladka-EtAl:2022:LREC,
  author   = {Hladka, Barbora and Mírovský, Jiří and Kopp, Matyáš and Moravec, Václav},
  title   = {Annotating Attribution in Czech News Server Articles},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1817--1823},
  abstract = {This paper focuses on detection of sources in the Czech articles published on a news server of Czech public radio. In particular, we search for attribution in sentences and we recognize attributed sources and their sentence context (signals). We organized a crowdsourcing annotation task that resulted in a data set of 2,167 stories with manually recognized signals and sources. In addition, the sources were classified into the classes of named and unnamed sources.},
  url   = {https://aclanthology.org/2022.lrec-1.193}
}

@InProceedings{gessler-EtAl:2022:LREC,

  author   = {Gessler, Luke and Schneider, Nathan and Ledford, Joseph C. and Blodgett, Austin},
  title   = {Xposition: An Online Multilingual Database of Adpositional Semantics},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1824--1830},
  abstract = {We present Xposition, an online platform for documenting adpositional semantics across languages in terms of supersenses (Schneider et al., 2018). More than just a lexical database, Xposition houses annotation guidelines, structured lexicographic documentation, and annotated corpora. Guidelines and documentation are stored as wiki pages for ease of editing, and described elements (supersenses, adpositions, etc.) are hyperlinked for ease of browsing. We describe how the platform structures information; its current contents across several languages; and aspects of the design of the web application that supports it, with special attention to how it supports datasets and standards that evolve over time.},
  url   = {https://aclanthology.org/2022.lrec-1.194}
}

@InProceedings{tracey-EtAl:2022:LREC1,
  author   = {Tracey, Jennifer and Bies, Ann and Getman, Jeremy and Griffitt, Kira and Strassel, Stephanie},
  title   = {A Study in Contradiction: Data and Annotation for AIDA Focusing on Informational Conflict in Russia-Ukraine Relations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1831--1838},
  abstract = {This paper describes data resources created for Phase 1 of the DARPA Active Interpretation of Disparate Alternatives (AIDA) program, which aims to develop language technology that can help humans manage large volumes of sometimes conflicting information to develop a comprehensive understanding of events around the world, even when such events are described in multiple media and languages. Especially important is the need for the technology to be capable of building multiple hypotheses to account for alternative interpretations of data imbued with informational conflict. The corpus described here is designed to support these goals. It focuses on the domain of Russia-Ukraine relations and contains multimedia source data in English, Russian and Ukrainian, annotated to support development and evaluation of systems that perform extraction of entities, events, and relations from individual multimedia documents, aggregate the information across

documents and languages, and produce multiple “hypotheses” about what has happened. This paper describes source data collection, annotation, and assessment.},
  url   = {https://aclanthology.org/2022.lrec-1.195}
}

@InProceedings{hadjmohamed-EtAl:2022:LREC,
  author   = {Hadj Mohamed, Najet and Ben Khelil, Cherifa and Savary, Agata and keskes, Iskandar and Antoine, Jean-Yves and Hadrich-Belguith, Lamia},
  title   = {Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1839--1848},
  abstract = {This paper describes our efforts to extend the PARSEME framework to Modern Standard Arabic. Theapplicability of the PARSEME guidelines was tested by measuring the inter-annotator agreement in theearly annotation stage. A subset of 1,062 sentences from the Prague Arabic Dependency Treebank PADTwas selected and annotated by two Arabic native speakers independently. Following their annotations, anew Arabic corpus with over 1,250 annotated VMWEs has been built. This corpus already exceeds thesmallest corpora of the PARSEME suite, and enables first observations. We discuss our annotation guide-line schema that shows full MWE annotation is realizable in Arabic where we get good inter-annotator agreement.},
  url   = {https://aclanthology.org/2022.lrec-1.196}
}

@InProceedings{figueroa-EtAl:2022:LREC,
  author   = {Figueroa, Carol and Adigwe, Adaeze and Ochs, Magalie and Skantze, Gabriel},
  title   = {Annotation of Communicative Functions of Short Feedback Tokens in Switchboard},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1849--1859},
  abstract = {There has been a lot of work on predicting the timing of feedback in conversational systems. However, there has been less focus on predicting the prosody and lexical form of feedback given their communicative function. Therefore, in this paper we present our preliminary annotations of the communicative functions of 1627 short feedback tokens from the Switchboard corpus and an analysis of their lexical realizations and prosodic characteristics. Since there is no standard scheme for annotating the communicative function of

feedback we propose our own annotation scheme. Although our work is ongoing, our preliminary analysis revealed lexical tokens such as "yeah" are ambiguous and therefore lexical forms alone are not indicative of the function. Both the lexical form and prosodic characteristics need to be taken into account in order to predict the communicative function. We also found that feedback functions have distinguishable prosodic characteristics in terms of duration, mean pitch, pitch slope, and pitch range.},
  url   = {https://aclanthology.org/2022.lrec-1.197}
}

@InProceedings{ajvazi-hardmeier:2022:LREC,
  author   = {Ajvazi, Adem and Hardmeier, Christian},
  title   = {A Dataset of Offensive Language in Kosovo Social Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1860--1869},
  abstract = {Social media are a central part of people's lives. Unfortunately, many public social media spaces are rife with bullying and offensive language, creating an unsafe environment for their users. In this paper, we present a new dataset for offensive language detection in Albanian. The dataset is composed of user-generated comments on Facebook and YouTube from the channels of selected Kosovo news platforms. It is annotated according to the three levels of the OLID annotation scheme. We also show results of a baseline system for offensive language classification based on a fine-tuned BERT model and compare with the Danish DKhate dataset, which is similar in scope and size. In a transfer learning setting, we find that merging the Albanian and Danish training sets leads to improved performance for prediction on Danish, but not Albanian, on both offensive language recognition and distinguishing targeted and untargeted offence.},
  url   = {https://aclanthology.org/2022.lrec-1.198}
}

@InProceedings{alhafni-habash-bouamor:2022:LREC,
  author   = {Alhafni, Bashar and Habash, Nizar and Bouamor, Houda},
  title   = {The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1870--1884},
  abstract = {Gender bias in natural language processing (NLP) applications, particularly machine translation, has been receiving

increasing attention. Much of the research on this issue has focused on mitigating gender bias in English NLP models and systems. Addressing the problem in poorly resourced, and/or morphologically rich languages has lagged behind, largely due to the lack of datasets and resources. In this paper, we introduce a new corpus for gender identification and rewriting in contexts involving one or two target users (I and/or You) – first and second grammatical persons with independent grammatical gender preferences. We focus on Arabic, a gender-marking morphologically rich language. The corpus has multiple parallel components: four combinations of 1st and 2nd person in feminine and masculine grammatical genders, as well as English, and English to Arabic machine translation output. This corpus expands on Habash et al. (2019)'s Arabic Parallel Gender Corpus (APGC v1.0) by adding second person targets as well as increasing the total number of sentences over 6.5 times, reaching over 590K words. Our new dataset will aid the research and development of gender identification, controlled text generation, and post-editing rewrite systems that could be used to personalize NLP applications and provide users with the correct outputs based on their grammatical gender preferences. We make the Arabic Parallel Gender Corpus (APGC v2.0) publicly available},
  url   = {https://aclanthology.org/2022.lrec-1.199}
}

@InProceedings{gladkoff-han:2022:LREC,
  author   = {Gladkoff, Serge and Han, Lifeng},
  title   = {HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {13--21},
  abstract = {Traditional automatic evaluation metrics for machine translation have been widely criticized by linguists due to their low accuracy, lack of transparency, focus on language mechanics rather than semantics, and low agreement with human quality evaluation. Human evaluations in the form of MQM-like scorecards have always been carried out in real industry setting by both clients and translation service providers (TSPs). However, traditional human translation quality evaluations are costly to perform and go into great linguistic detail, raise issues as to inter-rater reliability (IRR) and are not designed to measure quality of worse than premium quality translations. In this work, we introduce \textbf{HOPE}, a task-oriented and \textit{\textbf{h}}uman-centric evaluation framework for machine translation output based \textit{\textbf{o}}n professional \textit{\textbf{p}}ost-\textit{\textbf{e}}diting annotations. It contains only a limited number of commonly occurring error types, and uses a scoring model with geometric progression of error penalty points (EPPs) reflecting error severity level to each translation unit. The initial

experimental work carried out on English-Russian language pair MT outputs on marketing content type of text from highly technical domain reveals that our evaluation framework is quite effective in reflecting the MT output quality regarding both overall system-level performance and segment-level transparency, and it increases the IRR for error type interpretation. The approach has several key advantages, such as ability to measure and compare less than perfect MT output from different systems, ability to indicate human perception of quality, immediate estimation of the labor effort required to bring MT output to premium quality, low-cost and faster application, as well as higher IRR. Our experimental data is available at \url{https://github.com/lHan87/HOPE}.},
  url   = {https://aclanthology.org/2022.lrec-1.2}
}

@InProceedings{rosner-EtAl:2022:LREC,
  author   = {Rosner, Michael and Ahmadi, Sina and Apostol, Elena-Simona and Bosque-Gil, Julia and Chiarcos, Christian and Dojchinovski, Milan and Gkirtzou, Katerina and Gracia, Jorge and Gromann, Dagmar and Liebeskind, Chaya and Valūnaitė Oleškevičienė, Giedrė and Sérasset, Gilles and Truică, Ciprian-Octavian},
  title   = {Cross-Lingual Link Discovery for Under-Resourced Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {181--192},
  abstract = {In this paper, we provide an overview of current technologies for cross-lingual link discovery, and we discuss challenges, experiences and prospects of their application to under-resourced languages. We rst introduce the goals of cross-lingual linking and associated technologies, and in particular, the role that the Linked Data paradigm (Bizer et al., 2011) applied to language data can play in this context. We de ne under-resourced languages with a speci c focus on languages actively used on the internet, i.e., languages with a digitally versatile speaker community, but limited support in terms of language technology. We argue that languages for which considerable amounts of textual data and (at least) a bilingual word list are available, techniques for cross-lingual linking can be readily applied, and that these enable the implementation of downstream applications for under-resourced languages via the localisation and adaptation of existing technologies and resources.},
  url   = {https://aclanthology.org/2022.lrec-1.20}
}

@InProceedings{cheng-EtAl:2022:LREC1,
  author   = {Cheng, Daniel and Yan, Kyle and Keung, Phillip and Smith, Noah A.},
  title   = {The Engage Corpus: A Social Media Dataset for Text-

Based Recommender Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1885--1889},
  abstract = {Social media platforms play an increasingly important role as forums for public discourse. Many platforms use recommendation algorithms that funnel users to online groups with the goal of maximizing user engagement, which many commentators have pointed to as a source of polarization and misinformation. Understanding the role of NLP in recommender systems is an interesting research area, given the role that social media has played in world events. However, there are few standardized resources which researchers can use to build models that predict engagement with online groups on social media; each research group constructs datasets from scratch without releasing their version for reuse. In this work, we present a dataset drawn from posts and comments on the online message board Reddit. We develop baseline models for recommending subreddits to users, given the user's post and comment history. We also study the behavior of our recommender models on subreddits that were banned in June 2020 as part of Reddit's efforts to stop the dissemination of hate speech.},
  url   = {https://aclanthology.org/2022.lrec-1.200}
}

@InProceedings{rocha-EtAl:2022:LREC,
  author   = {Rocha, Gil and Trigo, Luís and Lopes Cardoso, Henrique and Sousa-Silva, Rui and Carvalho, Paula and Martins, Bruno and Won, Miguel},
  title   = {Annotating Arguments in a Corpus of Opinion Articles},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1890--1899},
  abstract = {Interest in argument mining has resulted in an increasing number of argument annotated corpora. However, most focus on English texts with explicit argumentative discourse markers, such as persuasive essays or legal documents. Conversely, we report on the first extensive and consolidated Portuguese argument annotation project focused on opinion articles. We briefly describe the annotation guidelines based on a multi-layered process and analyze the manual annotations produced, highlighting the main challenges of this textual genre. We then conduct a comprehensive inter-annotator agreement analysis, including argumentative discourse units, their classes and relations, and resulting graphs. This analysis reveals that each of these aspects tackles very different kinds of challenges. We observe differences in annotator profiles, motivating

our aim of producing a non-aggregated corpus containing the insights of every annotator. We note that the interpretation and identification of token-level arguments is challenging; nevertheless, tasks that focus on higher-level components of the argument structure can obtain considerable agreement. We lay down perspectives on corpus usage, exploiting its multi-faceted nature.},
  url   = {https://aclanthology.org/2022.lrec-1.201}
}

@InProceedings{abrami-EtAl:2022:LREC,
  author   = {Abrami, Giuseppe and Bagci, Mevlüt and Hammerla, Leon and Mehler, Alexander},
  title   = {German Parliamentary Corpus (GerParCor)},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1900--1906},
  abstract = {Parliamentary debates represent a large and partly unexploited treasure trove of publicly accessible texts. In the German-speaking area, there is a certain deficit of uniformly accessible and annotated corpora covering all German-speaking parliaments at the national and federal level. To address this gap, we introduce the German Parliamentary Corpus (GerParCor). GerParCor is a genre-specific corpus of (predominantly historical) German-language parliamentary protocols from three centuries and four countries, including state and federal level data. In addition, GerParCor contains conversions of scanned protocols and, in particular, of protocols in Fraktur converted via an OCR process based on Tesseract. All protocols were preprocessed by means of the NLP pipeline of spaCy3 and automatically annotated with metadata regarding their session date. GerParCor is made available in the XMI format of the UIMA project. In this way, GerParCor can be used as a large corpus of historical texts in the field of political communication for various tasks in NLP.},
  url   = {https://aclanthology.org/2022.lrec-1.202}
}

@InProceedings{novk-novk:2022:LREC,
  author   = {Novák, Attila and Novák, Borbála},
  title   = {NerKor+Cars-OntoNotes++},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1907--1916},
  abstract = {In this paper, we present an upgraded version of the Hungarian NYTK-NerKor named entity corpus, which contains about twice as many annotated spans and 7 times as many distinct entity types as the original version. We used an extended version of the

OntoNotes 5 annotation scheme including time and numerical expressions. NerKor is the newest and biggest NER corpus for Hungarian containing diverse domains. We applied cross-lingual transfer of NER models trained for other languages based on multilingual contextual language models to preannotate the corpus. We corrected the annotation semi-automatically and manually. Zero-shot preannotation was very effective with about 0.82 F1 score for the best model. We also added a 12000-token subcorpus on cars and other motor vehicles. We trained and release a transformer-based NER tagger for Hungarian using the annotation in the new corpus version, which provides similar performance to an identical model trained on the original version of the corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.203}
}

@InProceedings{burkhardt-EtAl:2022:LREC1,
  author   = {Burkhardt, Felix and Hacker, Anabell and Reichel, Uwe and Wierstorf, Hagen and Eyben, Florian and Schuller, Björn},
  title   = {A Comparative Cross Language View On Acted Databases Portraying Basic Emotions Utilising Machine Learning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1917--1924},
  abstract = {Since several decades emotional databases have been recorded by various laboratories. Many of them contain acted portrays of Darwin’s famous “big four” basic emotions. In this paper, we investigate in how far a selection of them are comparable by two approaches: on the one hand modeling similarity as performance in cross database machine learning experiments and on the other by analyzing a manually picked set of four acoustic features that represent different phonetic areas. It is interesting to see in how far specific databases (we added a synthetic one) perform well as a training set for others while some do not. Generally speaking, we found indications for both similarity as well as specificiality across languages.},
  url   = {https://aclanthology.org/2022.lrec-1.204}
}

@InProceedings{burkhardt-EtAl:2022:LREC2,
  author   = {Burkhardt, Felix and Wagner, Johannes and Wierstorf, Hagen and Eyben, Florian and Schuller, Björn},
  title   = {Nkululeko: A Tool For Rapid Speaker Characteristics Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {1925--1932},
  abstract = {We present advancements with a software tool called Nkululeko, that lets users perform (semi-) supervised machine learning experiments in the speaker characteristics domain. It is based on audformat, a format for speech database metadata description. Due to an interface based on configurable templates, it supports best practise and very fast setup of experiments without the need to be proficient in the underlying language: Python. The paper explains the handling of Nkululeko and presents two typical experiments: comparing the expert acoustic features with artificial neural net embeddings for emotion classification and speaker age regression.},
  url   = {https://aclanthology.org/2022.lrec-1.205}
}

@InProceedings{yu-EtAl:2022:LREC1,
  author   = {YU, Shi and Ponchard, Clara and Trouville, Roland and Hassid, Sergio and Demolin, Didier},
  title   = {Speech Aerodynamics Database, Tools and Visualisation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1933--1938},
  abstract = {Aerodynamic processes underlie the characteristics of the acoustic signal of speech sounds. The aerodynamics of speech give insights on acoustic outcome and help explain the mechanisms of speech production. This database was designed during an ARC project ”Dynamique des systèmes phonologiques” in which the study of aerodynamic constraints on speech production was an important target. Data were recorded between 1996 and 1999 at the Erasmus Hospital (Hôpital Erasme) of Université Libre de Bruxelles, Belgium and constitute one of the few datasets available on direct measurement of subglottal pressure and other aerodynamic parameters. The goal was to obtain a substantial amount of data with simultaneous recording, in various context, of the speech acoustic signal, subglottal pressure (Ps), intraoral pressure (Po), oral airflow (Qo) and nasal airflow (Qn). This database contains recordings of 2 English, 1 Amharic, and 7 French speakers and is provided with data conversion and visualisation tools. Another aim of this project was to obtain some reference values of the aerodynamics of speech production for female and male speakers uttering different types of segments and sentences in French.},
  url   = {https://aclanthology.org/2022.lrec-1.206}
}

@InProceedings{fougeron-EtAl:2022:LREC,
  author   = {Fougeron, Cécile and Audibert, Nicolas and Gendrot, cedric and Chardenon, Estelle and Wohmann, Louise},
  title   = {PATATRA and PATAFreq: two French databases for the documentation of within-speaker variability in speech},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1939--1944},
  abstract = {Our knowledge on speech is historically built on data comparing different speakers or data averaged across speakers. Consequently, little is known on the variability in the speech of a single individual. Experimental studies have shown that speakers adapt to the linguistic and the speaking contexts, and modify their speech according to their emotional or biological condition, etc. However, it is unclear how much speakers vary from one repetition to the next, and how comparable are recordings that are collected days, months or years apart. In this paper, we introduce two French databases which contain recordings of 9 to 11 speakers recorded over 9 to 18 sessions, allowing comparisons of speech tasks with a different delay between the repetitions: 3 repetitions within the same session, 6 to 10 repetitions on different days during a two months period, 5 to 9 repetitions on different years. Speakers are recorded on a large set of speech tasks including read and spontaneous speech as well as speech-like performance tasks. In this paper, we provide detailed descriptions of the two databases and available annotations. We conclude by an illustration on how these data can inform on within-speaker variability of speech.},
  url   = {https://aclanthology.org/2022.lrec-1.207}
}

@InProceedings{mukiibi-EtAl:2022:LREC,
  author   = {Mukiibi, Jonathan and Katumba, Andrew and Nakatumba-Nabende, Joyce and Hussein, Ali and Meyer, Joshua},
  title   = {The Makerere Radio Speech Corpus: A Luganda Radio Corpus for Automatic Speech Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1945--1954},
  abstract = {Building a usable radio monitoring automatic speech recognition (ASR) system is a challenging task for under-resourced languages and yet this is paramount in societies where radio is the main medium of public communication and discussions. Initial efforts by the United Nations in Uganda have proved how understanding the perceptions of rural people who are excluded from social media is important in national planning. However, these efforts are being challenged by the absence of transcribed speech datasets. In this paper, The Makerere Artificial Intelligence research lab releases a Luganda radio speech corpus of 155 hours. To our knowledge, this is the first publicly available radio dataset in sub-Saharan Africa. The paper describes the development of the voice corpus and presents baseline Luganda ASR performance results using Coqui STT toolkit, an

open-source speech recognition toolkit.},
  url   = {https://aclanthology.org/2022.lrec-1.208}
}

@InProceedings{rouvier-mohammadamini:2022:LREC,
  author   = {Rouvier, Mickael and Mohammadamini, Mohammad},
  title   = {Far-Field Speaker Recognition Benchmark Derived From The DiPCo Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1955--1959},
  abstract = {In this paper, we present a far-field speaker verification benchmark derived from the publicly-available DiPCo corpus. This corpus comprise three different tasks that involve enrollment and test conditions with single- and/or multi-channels recordings. The main goal of this corpus is to foster research in far-field and multi-channel text-independent speaker verification. Also, it can be used for other speaker recognition tasks such as dereverberation, denoising and speech enhancement. In addition, we release a Kaldi and SpeechBrain system to facilitate further research. And we validate the evaluation design with a single-microphone state-of-the-art speaker recognition system (i.e. ResNet-101). The results show that the proposed tasks are very challenging. And we hope these resources will inspire the speech community to develop new methods and systems for this challenging domain.},
  url   = {https://aclanthology.org/2022.lrec-1.209}
}

@InProceedings{dragos-EtAl:2022:LREC,
  author   = {Dragos, Valentina and Battistelli, Delphine and Etienne, Aline and Constable, Yolène},
  title   = {Angry or Sad ? Emotion Annotation for Extremist Content Characterisation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {193--201},
  abstract = {This paper examines the role of emotion annotations to characterize extremist content released on social platforms. The analysis of extremist content is important to identify user emotions towards some extremist ideas and to highlight the root cause of where emotions and extremist attitudes merge together. To address these issues our methodology combines knowledge from sociological and linguistic annotations to explore French extremist content collected online. For emotion linguistic analysis, the solution presented in this paper relies on a complex linguistic annotation

scheme. The scheme was used to annotate extremist text corpora in French. Data sets were collected online by following semi-automatic procedures for content selection and validation. The paper describes the integrated annotation scheme, the annotation protocol that was set-up for French corpora annotation and the results, e.g. agreement measures and remarks on annotation disagreements. The aim of this work is twofold: first, to provide a characterization of extremist contents; second, to validate the annotation scheme and to test its capacity to capture and describe various aspects of emotions.},
  url   = {https://aclanthology.org/2022.lrec-1.21}
}

@InProceedings{wang-gustafson-szkely:2022:LREC,
  author   = {Wang, Siyang and gustafson, joakim and Székely, Éva},
  title   = {Evaluating Sampling-based Filler Insertion with Spontaneous TTS},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1960--1969},
  abstract = {Inserting fillers (such as "um'', "like'') to clean speech text has a rich history of study. One major application is to make dialogue systems sound more spontaneous. The ambiguity of filler occurrence and inter-speaker difference make both modeling and evaluation difficult. In this paper, we study sampling-based filler insertion, a simple yet unexplored approach to inserting fillers. We propose an objective score called Filler Perplexity (FPP). We build three models trained on two single-speaker spontaneous corpora, and evaluate them with FPP and perceptual tests. We implement two innovations in perceptual tests, (1) evaluating filler insertion on dialogue systems output, (2) synthesizing speech with neural spontaneous TTS engines. FPP proves to be useful in analysis but does not correlate well with perceptual MOS. Perceptual results show little difference between compared filler insertion models including with ground-truth, which may be due to the ambiguity of what is good filler insertion and a strong neural spontaneous TTS that produces natural speech irrespective of input. Results also show preference for filler-inserted speech synthesized with spontaneous TTS. The same test using TTS based on read speech obtains the opposite results, which shows the importance of using spontaneous TTS in evaluating filler insertions. Audio samples: www.speech.kth.se/tts-demos/LREC22},
  url   = {https://aclanthology.org/2022.lrec-1.210}
}

@InProceedings{mihajlik-EtAl:2022:LREC,
  author   = {Mihajlik, Peter and Balog, Andras and Graczi, Tekla Etelka and Kohari, Anna and Tarján, Balázs and Mady, Katalin},
  title   = {BEA-Base: A Benchmark for ASR of Spontaneous

Hungarian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1970--1977},
  abstract = {Hungarian is spoken by 15 million people, still, easily accessible Automatic Speech Recognition (ASR) benchmark datasets – especially for spontaneous speech – have been practically unavailable. In this paper, we introduce BEA-Base, a subset of the BEA spoken Hungarian database comprising mostly spontaneous speech of 140 speakers. It is built specifically to assess ASR, primarily for conversational AI applications. After defining the speech recognition subsets and task, several baselines – including classic HMM-DNN hybrid and end-to-end approaches augmented by cross-language transfer learning – are developed using open-source toolkits. The best results obtained are based on multilingual self-supervised pretraining, achieving a 45\% recognition error rate reduction as compared to the classical approach – without the application of an external language model or additional supervised data. The results show the feasibility of using BEA-Base for training and evaluation of Hungarian speech recognition systems.},
  url   = {https://aclanthology.org/2022.lrec-1.211}
}

@InProceedings{barker-EtAl:2022:LREC,
  author   = {Barker, Emma and Barker, Jon and Gaizauskas, Robert and Ma, Ning and Paramita, Monica Lestari},
  title   = {SNuC: The Sheffield Numbers Spoken Language Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1978--1984},
  abstract = {We present SNuC, the first published corpus of spoken alphanumeric identifiers of the sort typically used as serial and part numbers in the manufacturing sector. The dataset contains recordings and transcriptions of over 50 native British English speakers, speaking over 13,000 multi-character alphanumeric sequences and totalling almost 20 hours of recorded speech. We describe requirements taken into account in the designing the corpus and the methodology used to construct it. We present summary statistics describing the corpus contents, as well as a preliminary investigation into errors in spoken alphanumeric identifiers. We validate the corpus by showing how it can be used to adapt a deep learning neural network based ASR system, resulting in improved recognition accuracy on the task of spoken alphanumeric identifier recognition. Finally, we discuss further potential uses for the corpus and for the tools developed to construct it.},
  url   = {https://aclanthology.org/2022.lrec-1.212}

}

@InProceedings{zhao-chodroff:2022:LREC,
  author   = {Zhao, Liang and Chodroff, Eleanor},
  title   = {The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1985--1990},
  abstract = {In the present paper, we introduce the ManDi Corpus, a spoken corpus of regional Mandarin dialects and Standard Mandarin. The corpus currently contains 357 recordings (about 9.6 hours) of monosyllabic words, disyllabic words, short sentences, a short passage and a poem, each produced in Standard Mandarin and in one of six regional Mandarin dialects: Beijing, Chengdu, Jinan, Taiyuan, Wuhan, and Xi'an Mandarin from 36 speakers. The corpus was collected remotely using participant-controlled smartphone recording apps. Word- and phone-level alignments were generated using Praat and the Montreal Forced Aligner. The pilot study of dialect-specific tone systems showed that with practicable design and decent recording quality, remotely collected speech data can be suitable for analysis of relative patterns in acoustic-phonetic realization. The corpus is available on OSF (https://osf.io/fgv4w/) for non-commercial use under a CC BY-NC 3.0 license.},
  url   = {https://aclanthology.org/2022.lrec-1.213}
}

@InProceedings{carbone-EtAl:2022:LREC,
  author   = {Carbone, Francesca and Bouchet, Gilles and Ghio, Alain and Legou, Thierry and André, Carine and lalain, muriel and Kadri, Sabrina and Petrone, Caterina and Procino, Federica and Giovanni, Antoine},
  title   = {The Speed-Vel Project: a Corpus of Acoustic and Aerodynamic Data to Measure Droplets Emission During Speech Interaction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {1991--1999},
  abstract = {Conversations (normal speech) or professional interactions (e.g., projected speech in the classroom) have been identified as situations with increased risk of exposure to SARS-CoV-2 due to the high production of droplets in the exhaled air. However, it is still unclear to what extent speech properties influence droplets emission during everyday life conversations. Here, we report the experimental protocol of three experiments aiming at measuring the velocity and the direction of the airflow,

the number and size of droplets spread during speech interactions in French. We consider different phonetic conditions, potentially leading to a modulation of speech droplets production, such as voice intensity (normal vs. loud voice), articulation manner of phonemes (type of consonants and vowels) and prosody (i.e., the melody of the speech). Findings from these experiments will allow future simulation studies to predict the transport, dispersion and evaporation of droplets emitted under different speech conditions.},
  url   = {https://aclanthology.org/2022.lrec-1.214}
}

@InProceedings{aicher-EtAl:2022:LREC3,
  author   = {Aicher, Annalena and Gazizullina, Alisa and Gusev, Aleksei and Matveev, Yuri and Minker, Wolfgang},
  title   = {Towards Speech-only Opinion-level Sentiment Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2000--2006},
  abstract = {The growing popularity of various forms of Spoken Dialogue Systems (SDS) raises the demand for their capability of implicitly assessing the speaker's sentiment from speech only. Mapping the latter on user preferences enables to adapt to the user and individualize the requested information while increasing user satisfaction. In this paper, we explore the integration of rank consistent ordinal regression into a speech-only sentiment prediction task performed by ResNet-like systems. Furthermore, we use speaker verification extractors trained on larger datasets as low-level feature extractors. An improvement of performance is shown by fusing sentiment and pre-extracted speaker embeddings reducing the speaker bias of sentiment predictions. Numerous experiments on Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) databases show that we beat the baselines of state-of-the-art unimodal approaches. Using speech as the only modality combined with optimizing an order-sensitive objective function gets significantly closer to the sentiment analysis results of state-of-the-art multimodal systems.},
  url   = {https://aclanthology.org/2022.lrec-1.215}
}

@InProceedings{vanboven-hirmer-conforti:2022:LREC,
  author   = {van Boven, Goya and Hirmer, Stephanie and Conforti, Costanza},
  title   = {At the Intersection of NLP and Sustainable Development: Exploring the Impact of Demographic-Aware Text Representations in Modeling Value on a Corpus of Interviews},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {2007--2021},
abstract = {This research explores automated text classification using data from Low- and Middle-Income Countries (LMICs). In particular, we explore enhancing text representations with demographic information of speakers in a privacy-preserving manner. We introduce the Demographic-Rich Qualitative UPV-Interviews Dataset (DR-QI), a rich dataset of qualitative interviews from rural communities in India and Uganda. The interviews were conducted following the latest standards for respectful interactions with illiterate speakers (Hirmer et al., 2021a). The interviews were later sentence-annotated for Automated User-Perceived Value (UPV) Classification (Conforti et al., 2020), a schema that classifies values expressed by speakers, resulting in a dataset of 5,333 sentences. We perform the UPV classification task, which consists of predicting which values are expressed in a given sentence, on the new DR-QI dataset. We implement a classification model using DistilBERT (Sanh et al., 2019), which we extend with demographic information. In order to preserve the privacy of speakers, we investigate encoding demographic information using autoencoders. We find that adding demographic information improves performance, even if such information is encoded. In addition, we find that the performance per UPV is linked to the number of occurrences of that value in our data.},
url = {https://aclanthology.org/2022.lrec-1.216}
}

@InProceedings{gref-EtAl:2022:LREC,
author = {Gref, Michael and Matthiesen, Nike and Hikkal Venugopala, Sreenivasa and Satheesh, Shalaka and Vijayananth, Aswinkumar and Ha, Duc Bach and Behnke, Sven and Köhler, Joachim},
title = {A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {2022--2031},
abstract = {For research in audiovisual interview archives often it is not only of interest what is said but also how. Sentiment analysis and emotion recognition can help capture, categorize and make these different facets searchable. In particular, for oral history archives, such indexing technologies can be of great interest. These technologies can help understand the role of emotions in historical remembering. However, humans often perceive sentiments and emotions ambiguously and subjectively. Moreover, oral history interviews have multi-layered levels of complex, sometimes contradictory, sometimes very subtle facets of emotions. Therefore, the question arises of the chance machines and humans have capturing

and assigning these into predefined categories. This paper investigates the ambiguity in human perception of emotions and sentiment in German oral history interviews and the impact on machine learning systems. Our experiments reveal substantial differences in human perception for different emotions. Furthermore, we report from ongoing machine learning experiments with different modalities. We show that the human perceptual ambiguity and other challenges, such as class imbalance and lack of training data, currently limit the opportunities of these technologies for oral history archives. Nonetheless, our work uncovers promising observations and possibilities for further research.},
  url   = {https://aclanthology.org/2022.lrec-1.217}
}

@InProceedings{cobeli-EtAl:2022:LREC,
  author   = {Cobeli, Ștefan and Iordache, Ioan-Bogdan and Yadav, Shweta and Caragea, Cornelia and Dinu, Liviu P. and Iliescu, Dragoș},
  title   = {Detecting Optimism in Tweets using Knowledge Distillation and Linguistic Analysis of Optimism},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2032--2041},
  abstract = {Finding the polarity of feelings in texts is a far-reaching task. Whilst the field of natural language processing has established sentiment analysis as an alluring problem, many feelings are left uncharted. In this study, we analyze the optimism and pessimism concepts from Twitter posts to effectively understand the broader dimension of psychological phenomenon. Towards this, we carried a systematic study by first exploring the linguistic peculiarities of optimism and pessimism in user-generated content. Later, we devised a multi-task knowledge distillation framework to simultaneously learn the target task of optimism detection with the help of the auxiliary task of sentiment analysis and hate speech detection. We evaluated the performance of our proposed approach on the benchmark Optimism/Pessimism Twitter dataset. Our extensive experiments show the superior- ity of our approach in correctly differentiating between optimistic and pessimistic users. Our human and automatic evaluation shows that sentiment analysis and hate speech detection are beneficial for optimism/pessimism detection.},
  url   = {https://aclanthology.org/2022.lrec-1.218}
}

@InProceedings{herath-EtAl:2022:LREC,
  author   = {Herath, Missaka and Chamindu, Kushan and Maduwantha, Hashan and Ranathunga, Surangika},
  title   = {Dataset and Baseline for Automatic Student Feedback Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2042--2049},
  abstract = {In this paper, we present a student feedback corpus, which contains 3000 instances of feedback written by university students. This dataset has been annotated for aspect terms, opinion terms, polarities of the opinion terms towards targeted aspects, document-level opinion polarities and sentence separations. We develop a hierarchical taxonomy for aspect categorization, which covers all the areas of the teaching-learning process. We annotated both implicit and explicit aspects using this taxonomy. Annotation methodology, difficulties faced during the annotation, and the details about the aspect term categorization have been discussed in detail. This annotated corpus can be used for Aspect Extraction, Aspect Level Sentiment Analysis, and Document Level Sentiment Analysis. Also the baseline results for all three tasks are given in the paper.},
  url   = {https://aclanthology.org/2022.lrec-1.219}
}

@InProceedings{zampieri-EtAl:2022:LREC,
  author   = {Zampieri, Nicolas and Ramisch, Carlos and Illina, Irina and Fohr, Dominique},
  title   = {Identification of Multiword Expressions in Tweets for Hate Speech Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {202--210},
  abstract = {Multiword expression (MWE) identification in tweets is a complex task due to the complex linguistic nature of MWEs combined with the non-standard language use in social networks. MWE features were shown to be helpful for hate speech detection (HSD). In this article, we present joint experiments on these two related tasks on English Twitter data: first we focus on the MWE identification task, and then we observe the influence of MWE-based features on the HSD task. For MWE identification, we compare the performance of two systems: lexicon-based and deep neural networks-based (DNN). We experimentally evaluate seven configurations of a state-of-the-art DNN system based on recurrent networks using pre-trained contextual embeddings from BERT. The DNN-based system outperforms the lexicon-based one thanks to its superior generalisation power, yielding much better recall. For the HSD task, we propose a new DNN architecture for incorporating MWE features. We confirm that MWE features are helpful for the HSD task. Moreover, the proposed DNN architecture beats previous MWE-based HSD systems by 0.4 to 1.1 F-measure points on average on four Twitter HSD corpora.},
  url   = {https://aclanthology.org/2022.lrec-1.22}

}

@InProceedings{tikhonov-EtAl:2022:LREC,
  author   = {Tikhonov, Alexey and Malkhasov, Alex and Manoshin, Andrey and Dima, George-Andrei and Cserháti, Réka and Hossain Asif, Md.Sadek and Sárdi, Matt},
  title   = {EENLP: Cross-lingual Eastern European NLP Index},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2050--2057},
  abstract = {Motivated by the sparsity of NLP resources for Eastern European languages, we present a broad index of existing Eastern European language resources (90+ datasets and 45+ models) published as a github repository open for updates from the community. Furthermore, to support the evaluation of commonsense reasoning tasks, we provide hand-crafted cross-lingual datasets for five different semantic tasks (namely news categorization, paraphrase detection, Natural Language Inference (NLI) task, tweet sentiment detection, and news sentiment detection) for some of the Eastern European languages. We perform several experiments with the existing multilingual models on these datasets to define the performance baselines and compare them to the existing results for other languages.},
  url   = {https://aclanthology.org/2022.lrec-1.220}
}

@InProceedings{agar-robnikikonja:2022:LREC,
  author   = {Žagar, Aleš and Robnik-Šikonja, Marko},
  title   = {Slovene SuperGLUE Benchmark: Translation and Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2058--2065},
  abstract = {We present SuperGLUE benchmark adapted and translated into Slovene using a combination of human and machine translation. We describe the translation process and problems arising due to differences in morphology and grammar. We evaluate the translated datasets in several modes: monolingual, cross-lingual, and multilingual, taking into account differences between machine and human translated training sets. The results show that the monolingual Slovene SloBERTa model is superior to massively multilingual and trilingual BERT models, but these also show a good cross-lingual performance on certain tasks. The performance of Slovene models still lags behind the best English models.},
  url   = {https://aclanthology.org/2022.lrec-1.221}
}

@InProceedings{zanonboito-EtAl:2022:LREC,
  author   = {Zanon Boito, Marcely and Bougares, Fethi and Barbier, Florentin and Gahbiche, Souhir and Barrault, Loïc and Rouvier, Mickael and Estève, Yannick},
  title   = {Speech Resources in the Tamasheq Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2066--2071},
  abstract = {In this paper we present two datasets for Tamasheq, a developing language mainly spoken in Mali and Niger. These two datasets were made available for the IWSLT 2022 low-resource speech translation track, and they consist of collections of radio recordings from daily broadcast news in Niger (Studio Kalangou) and Mali (Studio Tamani). We share (i) a massive amount of unlabeled audio data (671 hours) in five languages: French from Niger, Fulfulde, Hausa, Tamasheq and Zarma, and (ii) a smaller 17 hours parallel corpus of audio recordings in Tamasheq, with utterance-level translations in the French language. All this data is shared under the Creative Commons BY-NC-ND 3.0 license. We hope these resources will inspire the speech community to develop and benchmark models using the Tamasheq language.},
  url   = {https://aclanthology.org/2022.lrec-1.222}
}

@InProceedings{knyazeva-boulademareil-vernier:2022:LREC,
  author   = {knyazeva, elena and Boula de Mareüil, Philippe and Vernier, Frédéric},
  title   = {Aesop's fable "The North Wind and the Sun" Used as a Rosetta Stone to Extract and Map Spoken Words in Under-resourced Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2072--2079},
  abstract = {This paper describes a method of semi-automatic word spotting in minority languages, from one and the same Aesop fable "The North Wind and the Sun" translated in Romance languages/dialects from Hexagonal (i.e. Metropolitan) France and languages from French Polynesia. The first task consisted of finding out how a dozen words such as "wind" and "sun" were translated in over 200 versions collected in the field – taking advantage of orthographic similarity, word position and context. Occurrences of the translations were then extracted from the phone-aligned recordings. The results were judged accurate in 96–97\% of cases, both on the development corpus and a test set of unseen data. Corrected alignments were then mapped and basemaps were drawn to make various

linguistic phenomena immediately visible. The paper exemplifies how regular expressions may be used for this purpose. The final result, which takes the form of an online speaking atlas (enriching the https://atlas.limsi.fr website), enables us to illustrate lexical, morphological or phonetic variation.},
  url   = {https://aclanthology.org/2022.lrec-1.223}
}

@InProceedings{palenmichel-kim-lignos:2022:LREC,
  author   = {Palen-Michel, Chester and Kim, June and Lignos, Constantine},
  title   = {Multilingual Open Text Release 1: Public Domain News in 44 Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2080--2089},
  abstract = {We present a Multilingual Open Text (MOT), a new multilingual corpus containing text in 44 languages, many of which have limited existing text resources for natural language processing. The first release of the corpus contains over 2.8 million news articles and an additional 1 million short snippets (photo captions, video descriptions, etc.) published between 2001--2022 and collected from Voice of America's news websites. We describe our process for collecting, filtering, and processing the data. The source material is in the public domain, our collection is licensed using a creative commons license (CC BY 4.0), and all software used to create the corpus is released under the MIT License. The corpus will be regularly updated as additional documents are published.},
  url   = {https://aclanthology.org/2022.lrec-1.224}
}

@InProceedings{herrera-aich-parde:2022:LREC,
  author   = {Herrera, Megan and Aich, Ankit and Parde, Natalie},
  title   = {TweetTaglish: A Dataset for Investigating Tagalog-English Code-Switching},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2090--2097},
  abstract = {Deploying recent natural language processing innovations to low-resource settings allows for state-of-the-art research findings and applications to be accessed across cultural and linguistic borders. One low-resource setting of increasing interest is code-switching, the phenomenon of combining, swapping, or alternating the use of two or more languages in continuous

dialogue. In this paper, we introduce a large dataset (20k+ instances) to facilitate investigation of Tagalog-English code-switching, which has become a popular mode of discourse in Philippine culture. Tagalog is an Austronesian language and former official language of the Philippines spoken by over 23 million people worldwide, but it and Tagalog-English are under-represented in NLP research and practice. We describe our methods for data collection, as well as our labeling procedures. We analyze our resulting dataset, and finally conclude by providing results from a proof-of-concept regression task to establish dataset validity, achieving a strong performance benchmark (R2=0.797-0.909; RMSE=0.068-0.057).},
  url   = {https://aclanthology.org/2022.lrec-1.225}
}

@InProceedings{chiruzzo-EtAl:2022:LREC,
  author   = {Chiruzzo, Luis and Góngora, Santiago and Alvarez, Aldo and Giménez-Lugo, Gustavo and Agüero-Torales, Marvin and Rodríguez, Yliana},
  title   = {Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2098--2107},
  abstract = {This work presents a parallel corpus of Guarani-Spanish text aligned at sentence level. The corpus contains about 30,000 sentence pairs, and is structured as a collection of subsets from different sources, further split into training, development and test sets. A sample of sentences from the test set was manually annotated by native speakers in order to incorporate meta-linguistic annotations about the Guarani dialects present in the corpus and also the correctness of the alignment and translation. We also present some baseline MT experiments and analyze the results in terms of the subsets. We hope this corpus can be used as a benchmark for testing Guarani-Spanish MT systems, and aim to expand and improve the quality of the corpus in future iterations.},
  url   = {https://aclanthology.org/2022.lrec-1.226}
}

@InProceedings{vksna-EtAl:2022:LREC,
  author   = {Vīksna, Rinalds and Skadiņa, Inguna and Skadiņš, Raivis and Vasiļjevs, Andrejs and Rozis, Roberts},
  title   = {Assessing Multilinguality of Publicly Accessible Websites},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {2108--2116},
  abstract = {Although information on the Internet can be shared in many languages, the language presence on the World Wide Web is very disproportionate. The problem of multilingualism on the Web, in particular access, availability and quality of information in the world’s languages, has been the subject of UNESCO focus for several decades. Making European websites more multilingual is also one of the focal targets of the Connecting Europe Facility Automated Translation (CEF AT) digital service infrastructure. In order to monitor this goal, alongside other possible solutions, CEF AT needs a methodology and easy to use tool to assess the degree of multilingualism of a given website. In this paper we investigate methods and tools that automatically analyse the language diversity of the Web and propose indicators and methodology on how to measure the multilingualism of European websites. We also introduce a prototype tool based on open-source software that helps to assess multilingualism of the Web and can be independently run at set intervals. We also present initial results obtained with our tool that allows us to conclude that multilingualism on the Web is still a problem not only at the world level, but also at the European and regional level.},
  url   = {https://aclanthology.org/2022.lrec-1.227}
}

@InProceedings{kletz-EtAl:2022:LREC,
  author   = {Kletz, David and Langlais, Philippe and Lareau, François and Drouin, Patrick},
  title   = {A Methodology for Building a Diachronic Dataset of Semantic Shifts and its Application to QC-FR-Diac-V1.0, a Free Reference for French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2117--2125},
  abstract = {Different algorithms have been proposed to detect semantic shifts (changes in a word meaning over time) in a diachronic corpus. Yet, and somehow surprisingly, no reference corpus has been designed so far to evaluate them, leaving researchers to fallback to troublesome evaluation strategies. In this work, we introduce a methodology for the construction of a reference dataset for the evaluation of semantic shift detection, that is, a list of words where we know for sure whether they present a word meaning change over a period of interest. We leverage a state-of-the-art word-sense disambiguation model to associate a date of first appearance to all the senses of a word. Significant changes in sense distributions as well as clear stability are detected and the resulting words are inspected by experts using a dedicated interface before populating a reference dataset. As a proof of concept, we apply this methodology to a corpus of newspapers from Quebec covering the whole 20th century. We manually verified a subset of candidates, leading to QC-FR-Diac-V1.0, a corpus of 151

words allowing one to evaluate the identification of semantic shifts in French between 1910 and 1990.},
  url   = {https://aclanthology.org/2022.lrec-1.228}
}

@InProceedings{frohberg-binder:2022:LREC,
  author   = {Frohberg, Jörg and Binder, Frank},
  title   = {CRASS: A Novel Data Set and Benchmark to Test Counterfactual Reasoning of Large Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2126--2140},
  abstract = {We introduce the CRASS (counterfactual reasoning assessment) data set and benchmark utilizing questionized counterfactual conditionals as a novel and powerful tool to evaluate large language models. We present the data set design and benchmark. We test six state-of-the-art models against our benchmark. Our results show that it poses a valid challenge for these models and opens up considerable room for their improvement.},
  url   = {https://aclanthology.org/2022.lrec-1.229}
}

@InProceedings{jantscher-kern:2022:LREC,
  author   = {Jantscher, Michael and Kern, Roman},
  title   = {Causal Investigation of Public Opinion during the COVID-19 Pandemic via Social Media Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {211--226},
  abstract = {Understanding the needs and fears of citizens, especially during a pandemic such as COVID-19, is essential for any government or legislative entity. An effective COVID-19 strategy further requires that the public understand and accept the restriction plans imposed by these entities. In this paper, we explore a causal mediation scenario in which we want to emphasize the use of NLP methods in combination with methods from economics and social sciences. Based on sentiment analysis of Tweets towards the current COVID-19 situation in the UK and Sweden, we conduct several causal inference experiments and attempt to decouple the effect of government restrictions on mobility behavior from the effect that occurs due to public perception of the COVID-19 strategy in a country. To avoid biased results we control for valid country specific epidemiological and time-varying confounders. Comprehensive experiments show that not all changes in mobility are caused by countries implemented policies but also by the support of individuals in the fight against this pandemic. We find that social

media texts are an important source to capture citizens' concerns and trust in policy makers and are suitable to evaluate the success of government policies.},
  url   = {https://aclanthology.org/2022.lrec-1.23}
}

@InProceedings{costajuss-basta-gllego:2022:LREC,
  author   = {Costa-jussà, Marta R. and Basta, Christine and Gállego, Gerard I.},
  title   = {Evaluating Gender Bias in Speech Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2141--2147},
  abstract = {The scientific community is increasingly aware of the necessity to embrace pluralism and consistently represent major and minor social groups. Currently, there are no standard evaluation techniques for different types of biases. Accordingly, there is an urgent need to provide evaluation sets and protocols to measure existing biases in our automatic systems. Evaluating the biases should be an essential step towards mitigating them in the systems. This paper introduces WinoST, a new freely available challenge set for evaluating gender bias in speech translation. WinoST is the speech version of WinoMT, an MT challenge set, and both follow an evaluation protocol to measure gender accuracy. Using an S-Transformer end-to-end speech translation system, we report the gender bias evaluation on four language pairs, and we reveal the inaccuracies in translations generating gender-stereotyped translations.},
  url   = {https://aclanthology.org/2022.lrec-1.230}
}

@InProceedings{scholman-EtAl:2022:LREC1,
  author   = {Scholman, Merel and Pyatkin, Valentina and Yung, Frances and Dagan, Ido and Tsarfaty, Reut and Demberg, Vera},
  title   = {Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2148--2156},
  abstract = {Obtaining linguistic annotation from novice crowdworkers is far from trivial. A case in point is the annotation of discourse relations, which is a complicated task. Recent methods have obtained promising results by extracting relation labels from either discourse connectives (DCs) or question-answer (QA) pairs that participants provide. The current contribution studies the effect of worker selection and training on the agreement on implicit

relation labels between workers and gold labels, for both the DC and the QA method. In Study 1, workers were not specifically selected or trained, and the results show that there is much room for improvement. Study 2 shows that a combination of selection and training does lead to improved results, but the method is cost- and time-intensive. Study 3 shows that a selection-only approach is a viable alternative; it results in annotations of comparable quality compared to annotations from trained participants. The results generalized over both the DC and QA method and therefore indicate that a selection-only approach could also be effective for other crowdsourced discourse annotation tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.231}
}

@InProceedings{kulkarni-EtAl:2022:LREC,
  author   = {Kulkarni, Hrishikesh and MacAvaney, Sean and Goharian, Nazli and Frieder, Ophir},
  title   = {TBD3: A Thresholding-Based Dynamic Depression Detection from Social Media for Low-Resource Users},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2157--2165},
  abstract = {Social media are heavily used by many users to share their mental health concerns and diagnoses. This trend has turned social media into a large-scale resource for researchers focused on detecting mental health conditions. Social media usage varies considerably across individuals. Thus, classification of patterns, including detecting signs of depression, must account for such variation. We address the disparity in classification effectiveness for users with little activity (e.g., new users). Our evaluation, performed on a large-scale dataset, shows considerable detection discrepancy based on user posting frequency. For instance, the F1 detection score of users with an above-median versus below-median number of posts is greater than double (0.803 vs 0.365) using a conventional CNN-based model; similar results were observed on lexical and transformer-based classifiers. To complement this evaluation, we propose a dynamic thresholding technique that adjusts the classifier's sensitivity as a function of the number of posts a user has. This technique alone reduces the margin between users with many and few posts, on average, by 45\% across all methods and increases overall performance, on average, by 33\%. These findings emphasize the importance of evaluating and tuning natural language systems for potentially vulnerable populations.},
  url   = {https://aclanthology.org/2022.lrec-1.232}
}

@InProceedings{ghosh-EtAl:2022:LREC,
  author   = {Ghosh, Sayontan and Singh, Amanpreet and Merenstein, Alex and Su, Wei and Smolka, Scott A. and Zadok, Erez and Balasubramanian, Niranjan},

  title   = {SpecNFS: A Challenge Dataset Towards Extracting Formal Models from Natural Language Specifications},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2166--2176},
  abstract = {Can NLP assist in building formal models for verifying complex systems? We study this challenge in the context of parsing Network File System (NFS) specifications. We define a semantic-dependency problem over SpecIR, a representation language we introduce to model sentences appearing in NFS specification documents (RFCs) as IF-THEN statements, and present an annotated dataset of 1,198 sentences. We develop and evaluate semantic-dependency parsing systems for this problem. Evaluations show that even when using a state-of-the-art language model, there is significant room for improvement, with the best models achieving an F1 score of only 60.5 and 33.3 in the named-entity-recognition and dependency-link-prediction sub-tasks, respectively. We also release additional unlabeled data and other domain-related texts. Experiments show that these additional resources increase the F1 measure when used for simple domain-adaption and transfer-learning-based approaches, suggesting fruitful directions for further research},
  url   = {https://aclanthology.org/2022.lrec-1.233}
}

@InProceedings{bai-stede:2022:LREC,
  author   = {Bai, Xiaoyu and Stede, Manfred},
  title   = {Argument Similarity Assessment in German for Intelligent Tutoring: Crowdsourced Dataset and First Experiments},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2177--2187},
  abstract = {NLP technologies such as text similarity assessment, question answering and text classification are increasingly being used to develop intelligent educational applications. The long-term goal of our work is an intelligent tutoring system for German secondary schools, which will support students in a school exercise that requires them to identify arguments in an argumentative source text. The present paper presents our work on a central subtask, viz. the automatic assessment of similarity between a pair of argumentative text snippets in German. In the designated use case, students write out key arguments from a given source text; the tutoring system then evaluates them against a target reference, assessing the similarity level between student work and the reference. We collect a dataset for our similarity assessment task through crowdsourcing as authentic German student data are scarce;

we label the collected text pairs with similarity scores on a 5-point scale and run first experiments on the task. We see that a model based on BERT shows promising results, while we also discuss some challenges that we observe.},
  url   = {https://aclanthology.org/2022.lrec-1.234}
}

@InProceedings{jain-EtAl:2022:LREC,
  author   = {Jain, Nishtha and Groves, Declan and Specia, Lucia and Popović, Maja},
  title   = {Leveraging Pre-trained Language Models for Gender Debiasing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2188--2195},
  abstract = {Studying and mitigating gender and other biases in natural language have become important areas of research from both algorithmic and data perspectives. This paper explores the idea of reducing gender bias in a language generation context by generating gender variants of sentences. Previous work in this field has either been rule-based or required large amounts of gender balanced training data. These approaches are however not scalable across multiple languages, as creating data or rules for each language is costly and time-consuming. This work explores a light-weight method to generate gender variants for a given text using pre-trained language models as the resource, without any task-specific labelled data. The approach is designed to work on multiple languages with minimal changes in the form of heuristics. To showcase that, we have tested it on a high-resourced language, namely Spanish, and a low-resourced language from a different family, namely Serbian. The approach proved to work very well on Spanish, and while the results were less positive for Serbian, it showed potential even for languages where pre-trained models are less effective.},
  url   = {https://aclanthology.org/2022.lrec-1.235}
}

@InProceedings{delapeasarracn-rosso:2022:LREC,
  author   = {De la Peña Sarracén, Gretel Liz and Rosso, Paolo},
  title   = {Unsupervised Embeddings with Graph Auto-Encoders for Multi-domain and Multilingual Hate Speech Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2196--2204},
  abstract = {Hate speech detection is a prominent and challenging task, since hate messages are often expressed in subtle ways and with characteristics that may vary depending on the author. Hence,

many models suffer from the generalization problem. However, retrieving and monitoring hateful content on social media is a current necessity. In this paper, we propose an unsupervised approach using Graph Auto-Encoders (GAE), which allows us to avoid using labeled data when training the representation of the texts. Specifically, we represent texts as nodes of a graph, and use a transformer layer together with a convolutional layer to encode these nodes in a low-dimensional space. As a result, we obtain embeddings that can be decoded into a reconstruction of the original network. Our main idea is to learn a model with a set of texts without supervision, in order to generate embeddings for the nodes: nodes with the same label should be close in the embedding space, which, in turn, should allow us to distinguish among classes. We employ this strategy to detect hate speech in multi-domain and multilingual sets of texts, where our method shows competitive results on small datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.236}
}

@InProceedings{heinrich-viaud-belblidia:2022:LREC,
  author   = {Heinrich, Quentin and Viaud, Gautier and
Belblidia, Wacim},
  title   = {FQuAD2.0: French Question Answering and Learning When
You Don't Know},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2205--2214},
  abstract = {Question Answering, including Reading Comprehension,
is one of the NLP research areas that has seen significant
scientific breakthroughs over the past few years, thanks to the
concomitant advances in Language Modeling. Most of these
breakthroughs, however, are centered on the English language. In
2020, as a first strong initiative to bridge the gap to the French
language, Illuin Technology introduced FQuAD1.1, a French Native
Reading Comprehension dataset composed of 60,000+ questions and
answers samples extracted from Wikipedia articles. Nonetheless,
Question Answering models trained on this dataset have a major
drawback: they are not able to predict when a given question has no
answer in the paragraph of interest, therefore making unreliable
predictions in various industrial use-cases. We introduce FQuAD2.0,
which extends FQuAD with 17,000+ unanswerable questions, annotated
adversarially, in order to be similar to answerable ones. This new
dataset, comprising a total of almost 80,000 questions, makes it
possible to train French Question Answering models with the ability
of distinguishing unanswerable questions from answerable ones. We
benchmark several models with this dataset: our best model, a fine-
tuned CamemBERT-large, achieves a F1 score of 82.3\% on this
classification task, and a F1 score of 83\% on the Reading
Comprehension task.},
  url   = {https://aclanthology.org/2022.lrec-1.237}

}

@InProceedings{toraman-ahinu-yilmaz:2022:LREC,
  author   = {Toraman, Cagri and Şahinuç, Furkan and Yilmaz, Eyup},
  title   = {Large-Scale Hate Speech Detection with Cross-Domain Transfer},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2215--2225},
  abstract = {The performance of hate speech detection models relies on the datasets on which the models are trained. Existing datasets are mostly prepared with a limited number of instances or hate domains that define hate topics. This hinders large-scale analysis and transfer learning with respect to hate domains. In this study, we construct large-scale tweet datasets for hate speech detection in English and a low-resource language, Turkish, consisting of human-labeled 100k tweets per each. Our datasets are designed to have equal number of tweets distributed over five domains. The experimental results supported by statistical tests show that Transformer-based language models outperform conventional bag-of-words and neural models by at least 5\% in English and 10\% in Turkish for large-scale hate speech detection. The performance is also scalable to different training sizes, such that 98\% of performance in English, and 97\% in Turkish, are recovered when 20\% of training instances are used. We further examine the generalization ability of cross-domain transfer among hate domains. We show that 96\% of the performance of a target domain in average is recovered by other domains for English, and 92\% for Turkish. Gender and religion are more successful to generalize to other domains, while sports fail most.},
  url   = {https://aclanthology.org/2022.lrec-1.238}
}

@InProceedings{meyer-elsweiler:2022:LREC,
  author   = {Meyer, Selina and Elsweiler, David},
  title   = {GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2226--2235},
  abstract = {Health behaviour change is a difficult and prolonged process that requires sustained motivation and determination. Conversa- tional agents have shown promise in supporting the change process in the past. One therapy approach that facilitates change and has been used as a framework for conversational agents is

motivational interviewing. However, existing implementations of this therapy approach lack the deep understanding of user utterances that is essential to the spirit of motivational interviewing. To address this lack of understanding, we introduce the GLoHBCD, a German dataset of naturalistic language around health behaviour change. Data was sourced from a popular German weight loss forum and annotated using theoretically grounded motivational interviewing categories. We describe the process of dataset construction and present evaluation results. Initial experiments suggest a potential for broad applicability of the data and the resulting classifiers across different behaviour change domains. We make code to replicate the dataset and experiments available on Github.},
  url   = {https://aclanthology.org/2022.lrec-1.239}
}

@InProceedings{nakwijit-purver:2022:LREC,
  author   = {Nakwijit, Pakawat and Purver, Matthew},
  title   = {Misspelling Semantics in Thai},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {227--236},
  abstract = {User-generated content is full of misspellings. Rather than being just random noise, we hypothesise that many misspellings contain hidden semantics that can be leveraged for language understanding tasks. This paper presents a fine-grained annotated corpus of misspelling in Thai, together with an analysis of misspelling intention and its possible semantics to get a better understanding of the misspelling patterns observed in the corpus. In addition, we introduce two approaches to incorporate the semantics of misspelling: Misspelling Average Embedding (MAE) and Misspelling Semantic Tokens (MST). Experiments on a sentiment analysis task confirm our overall hypothesis: additional semantics from misspelling can boost the micro F1 score up to 0.4-2\%, while blindly normalising misspelling is harmful and suboptimal.},
  url   = {https://aclanthology.org/2022.lrec-1.24}
}

@InProceedings{hendrickx:2022:LREC,
  author   = {Hendrickx, Iris},
  title   = {Creating a Data Set of Abstractive Summaries of Turn-labeled Spoken Human-Computer Conversations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2236--2244},
  abstract = {Digital recorded written and spoken dialogues are becoming increasingly available as an effect of the technological

advances such as online messenger services and the use of chatbots. Summaries are a natural way of presenting the important information gathered from dialogues. We present a unique data set that consists of Dutch spoken human-computer conversations, an annotation layer of turn labels, and conversational abstractive summaries of user answers. The data set is publicly available for research purposes.},
  url   = {https://aclanthology.org/2022.lrec-1.240}
}

@InProceedings{cui-EtAl:2022:LREC1,
  author   = {Cui, Wen and Rolston, Leanne and Walker, Marilyn and Hockey, Beth Ann},
  title   = {OpenEL: An Annotated Corpus for Entity Linking and Discourse in Open Domain Dialogue},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2245--2256},
  abstract = {Entity linking in dialogue is the task of mapping entity mentions in utterances to a target knowledge base. Prior work on entity linking has mainly focused on well-written articles such as Wikipedia, annotated newswire, or domain-specific datasets. We extend the study of entity linking to open domain dialogue by presenting the OpenEL corpus: an annotated multi-domain corpus for linking entities in natural conversation to Wikidata. Each dialogic utterance in 179 dialogues over 12 topics from the EDINA dataset has been annotated for entities realized by definite referring expressions as well as anaphoric forms such as he, she, it and they. This dataset supports training and evaluation of entity linking in open-domain dialogue, as well as analysis of the effect of using dialogue context and anaphora resolution in model training. It could also be used for fine-tuning a coreference resolution algorithm. To the best of our knowledge, this is the first substantial entity linking corpus publicly available for open-domain dialogue. We also establish baselines for this task using several existing entity linking systems. We found that the Transformer-based system Flair + BLINK has the best performance with a 0.65 F1 score. Our results show that dialogue context is extremely beneficial for entity linking in conversations, with Flair + Blink achieving an F1 of 0.61 without discourse context. These results also demonstrate the remaining performance gap between the baselines and human performance, highlighting the challenges of entity linking in open-domain dialogue, and suggesting many avenues for future research using OpenEL.},
  url   = {https://aclanthology.org/2022.lrec-1.241}
}

@InProceedings{willemsen-kalpakchi-skantze:2022:LREC,
  author   = {Willemsen, Bram and Kalpakchi, Dmytro and Skantze, Gabriel},
  title   = {Collecting Visually-Grounded Dialogue with A Game Of

Sorts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2257--2268},
  abstract = {An idealized, though simplistic, view of the referring expression production and grounding process in (situated) dialogue assumes that a speaker must merely appropriately specify their expression so that the target referent may be successfully identified by the addressee. However, referring in conversation is a collaborative process that cannot be aptly characterized as an exchange of minimally-specified referring expressions. Concerns have been raised regarding assumptions made by prior work on visually-grounded dialogue that reveal an oversimplified view of conversation and the referential process. We address these concerns by introducing a collaborative image ranking task, a grounded agreement game we call "A Game Of Sorts". In our game, players are tasked with reaching agreement on how to rank a set of images given some sorting criterion through a largely unrestricted, role-symmetric dialogue. By putting emphasis on the argumentation in this mixed-initiative interaction, we collect discussions that involve the collaborative referential process. We describe results of a small-scale data collection experiment with the proposed task. All discussed materials, which includes the collected data, the codebase, and a containerized version of the application, are publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.242}
}

@InProceedings{hoefels-ltekin-mdroane:2022:LREC,
  author   = {Hoefels, Diana Constantina and Çöltekin, Çağrı and Mădroane, Irina Diana},
  title   = {CoRoSeOf - An Annotated Corpus of Romanian Sexist and Offensive Tweets},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2269--2281},
  abstract = {This paper introduces CoRoSeOf, a large corpus of Romanian social media manually annotated for sexist and offensive language. We describe the annotation process of the corpus, provide initial analyses, and baseline classification results for sexism detection on this data set. The resulting corpus contains 39 245 tweets, annotated by multiple annotators (with an agreement rate of Fleiss’κ= 0.45), following the sexist label set of a recent study. The automatic sexism detection yields scores similar to some of the earlier studies (macro averaged F1 score of 83.07\% on binary classification task). We release the corpus with a permissive license.},

  url   = {https://aclanthology.org/2022.lrec-1.243}
}

@InProceedings{almanea-poesio:2022:LREC,
  author   = {Almanea, Dina and Poesio, Massimo},
  title   = {ArMIS - The Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2282--2291},
  abstract = {The use of misogynistic and sexist language has increased in recent years in social media, and is increasing in the Arabic world in reaction to reforms attempting to remove restrictions on women lives. However, there are few benchmarks for Arabic misogyny and sexism detection, and in those the annotations are in aggregated form even though misogyny and sexism judgments are found to be highly subjective. In this paper we introduce an Arabic misogyny and sexism dataset (ArMIS) characterized by providing annotations from annotators with different degree of religious beliefs, and provide evidence that such differences do result in disagreements. To the best of our knowledge, this is the first dataset to study in detail the effect of beliefs on misogyny and sexism annotation. We also discuss proof-of-concept experiments showing that a dataset in which disagreements have not been reconciled can be used to train state-of-the-art models for misogyny and sexism detection; and consider different ways in which such models could be evaluated.},
  url   = {https://aclanthology.org/2022.lrec-1.244}
}

@InProceedings{yang-achard-pelachaud:2022:LREC,
  author   = {YANG, Liu and ACHARD, Catherine and PELACHAUD, Catherine},
  title   = {Annotating Interruption in Dyadic Human Interaction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2292--2297},
  abstract = {Integrating the existing interruption and turn switch classification methods, we propose a new annotation schema to annotate different types of interruptions through timeliness, switch accomplishment and speech content level. The proposed method is able to distinguish smooth turn exchange, backchannel and interruption (including interruption types) and to annotate dyadic conversation. We annotated the French part of NoXi corpus with the proposed structure and use these annotations to study the probability distribution and duration of each turn switch type.},

  url   = {https://aclanthology.org/2022.lrec-1.245}
}

@InProceedings{tan-EtAl:2022:LREC,
  author   = {Tan, Fiona Anting and Hürriyetoğlu, Ali and
Caselli, Tommaso and Oostdijk, Nelleke and Nomoto, Tadashi and
Hettiarachchi, Hansi and Ameer, Iqra and Uca, Onur and Liza,
Farhana Ferdousi and Hu, Tiancheng},
  title   = {The Causal News Corpus: Annotating Causal Relations
in Event Sentences from News},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2298--2310},
  abstract = {Despite the importance of understanding causality,
corpora addressing causal relations are limited. There is a
discrepancy between existing annotation guidelines of event
causality and conventional causality corpora that focus more on
linguistics. Many guidelines restrict themselves to include only
explicit relations or clause-based arguments. Therefore, we propose
an annotation schema for event causality that addresses these
concerns. We annotated 3,559 event sentences from protest event news
with labels on whether it contains causal relations or not. Our
corpus is known as the Causal News Corpus (CNC). A neural network
built upon a state-of-the-art pre-trained language model performed
well with 81.20\% F1 score on test set, and 83.46\% in 5-folds
cross-validation. CNC is transferable across two external corpora:
CausalTimeBank (CTB) and Penn Discourse Treebank (PDTB). Leveraging
each of these external datasets for training, we achieved up to
approximately 64\% F1 on the CNC test set without additional fine-
tuning. CNC also served as an effective training and pre-training
dataset for the two external corpora. Lastly, we demonstrate the
difficulty of our task to the layman in a crowd-sourced annotation
exercise. Our annotated corpus is publicly available, providing a
valuable resource for causal text mining researchers.},
  url   = {https://aclanthology.org/2022.lrec-1.246}
}

@InProceedings{hedstrm-EtAl:2022:LREC,
  author   = {Hedström, Staffan and Mollberg, David Erik and
Þórhallsdóttir, Ragnheiður and Guðnason, Jón},
  title   = {Samrómur: Crowd-sourcing large amounts of data},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2311--2316},
  abstract = {This contribution describes the collection of a large
and diverse corpus for speech recognition and similar tools using

crowd-sourced donations. We have built a collection platform inspired by Mozilla Common Voice and specialized it to our needs. We discuss the importance of engaging the community and motivating it to contribute, in our case through competitions. Given the incentive and a platform to easily read in large amounts of utterances, we have observed four cases of speakers freely donating over 10 thousand utterances. We have also seen that women are keener to participate in these events throughout all age groups. Manually verifying a large corpus is a monumental task and we attempt to automatically verify parts of the data using tools like Marosijo and the Montreal Forced Aligner. The method proved helpful, especially for detecting invalid utterances and halving the work needed from crowd-sourced verification.},
  url   = {https://aclanthology.org/2022.lrec-1.247}
}

@InProceedings{roller-EtAl:2022:LREC,
  author   = {Roller, Roland and Burchardt, Aljoscha and Feldhus, Nils and Seiffe, Laura and Budde, Klemens and Ronicke, Simon and Osmanodja, Bilgin},
  title   = {An Annotated Corpus of Textual Explanations for Clinical Decision Support},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2317--2326},
  abstract = {In recent years, machine learning for clinical decision support has gained more and more attention. In order to introduce such applications into clinical practice, a good performance might be essential, however, the aspect of trust should not be underestimated. For the treating physician using such a system and being (legally) responsible for the decision made, it is particularly important to understand the system's recommendation. To provide insights into a model's decision, various techniques from the field of explainability (XAI) have been proposed whose output is often enough not targeted to the domain experts that want to use the model. To close this gap, in this work, we explore how explanations could possibly look like in future. To this end, this work presents a dataset of textual explanations in context of decision support. Within a reader study, human physicians estimated the likelihood of possible negative patient outcomes in the near future and justified each decision with a few sentences. Using those sentences, we created a novel corpus, annotated with different semantic layers. Moreover, we provide an analysis of how those explanations are constructed, and how they change depending on physician, on the estimated risk and also in comparison to an automatic clinical decision support system with feature importance.},
  url   = {https://aclanthology.org/2022.lrec-1.248}
}

@InProceedings{passali-EtAl:2022:LREC,

  author   = {Passali, Tatiana and Mavropoulos, Thanassis and Tsoumakas, Grigorios and Meditskos, Georgios and Vrochidis, Stefanos},
  title   = {LARD: Large-scale Artificial Disfluency Generation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2327--2336},
  abstract = {Disfluency detection is a critical task in real-time dialogue systems. However, despite its importance, it remains a relatively unexplored field, mainly due to the lack of appropriate datasets. At the same time, existing datasets suffer from various issues, including class imbalance issues, which can significantly affect the performance of the model on rare classes, as it is demonstrated in this paper. To this end, we propose LARD, a method for generating complex and realistic artificial disfluencies with little effort. The proposed method can handle three of the most common types of disfluencies: repetitions, replacements, and restarts. In addition, we release a new large-scale dataset with disfluencies that can be used on four different tasks: disfluency detection, classification, extraction, and correction. Experimental results on the LARD dataset demonstrate that the data produced by the proposed method can be effectively used for detecting and removing disfluencies, while also addressing limitations of existing datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.249}
}

@InProceedings{moriceau-benamara-boumadane:2022:LREC,
  author   = {MORICEAU, Véronique and Benamara, Farah and Boumadane, Abdelmoumene},
  title   = {Automatic Detection of Stigmatizing Uses of Psychiatric Terms on Twitter},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {237--243},
  abstract = {Psychiatry and people suffering from mental disorders have often been given a pejorative label that induces social rejection.Many studies have addressed discourse content about psychiatry on social media, suggesting that they convey stigmatizingrepresentations of mental health disorders. In this paper, we focus for the first time on the use of psychiatric terms in tweetsin French. We first describe the annotated dataset that we use. Then we propose several deep learning models to detectautomatically (1) the different types of use of psychiatric terms (medical use, misuse or irrelevant use), and (2) the polarityof the tweet. We show that polarity detection can be

improved when done in a multitask framework in combination with typeof use detection. This confirms the observations made manually on several datasets, namely that the polarity of a tweet iscorrelated to the type of term use (misuses are mostly negative whereas medical uses are neutral). The results are interesting forboth tasks and it allows to consider the possibility for performant automatic approaches in order to conduct real-time surveyson social media, larger and less expensive than existing manual ones},
  url   = {https://aclanthology.org/2022.lrec-1.25}
}

@InProceedings{jiang-EtAl:2022:LREC1,
  author   = {Jiang, Yuru and Xu, Yang and Zhan, Yuhang and He, Weikai and Wang, Yilin and Xi, Zixuan and Wang, Meiyun and Li, Xinyu and Li, Yu and Yu, Yanchao},
  title   = {The CRECIL Corpus: a New Dataset for Extraction of Relations between Characters in Chinese Multi-party Dialogues},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2337--2344},
  abstract = {We describe a new freely available Chinese multi-party dialogue dataset for automatic extraction of dialogue-based character relationships. The data has been extracted from the original TV scripts of a Chinese sitcom called ``I Love My Home" with complex family-based human daily spoken conversations in Chinese. First, we introduced human annotation scheme for both global Character relationship map and character reference relationship. And then we generated the dialogue-based character relationship triples. The corpus annotates relationships between 140 entities in total. We also carried out a data exploration experiment by deploying a BERT-based model to extract character relationships on the CRECIL corpus and another existing relation extraction corpus (DialogRE \cite{yu2020dialogue}).The results demonstrate that extracting character relationships is more challenging in CRECIL than in DialogRE.},
  url   = {https://aclanthology.org/2022.lrec-1.250}
}

@InProceedings{abdulrahim-EtAl:2022:LREC,
  author   = {Abdulrahim, Dana and Inoue, Go and Shamsan, Latifa and Khalifa, Salam and Habash, Nizar},
  title   = {The Bahrain Corpus: A Multi-genre Corpus of Bahraini Arabic},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {2345--2352},
  abstract = {In recent years, the focus on developing natural language processing (NLP) tools for Arabic has shifted from Modern Standard Arabic to various Arabic dialects. Various corpora of various sizes and representing different genres, have been created for a number of Arabic dialects. As far as Gulf Arabic is concerned, Gumar Corpus (Khalifa et al., 2016) is the largest corpus, to date, that includes data representing the dialectal Arabic of the six Gulf Cooperation Council countries (Bahrain, Kuwait, Saudi Arabia, Qatar, United Arab Emirates, and Oman), particularly in the genre of "online forum novels". In this paper, we present the Bahrain Corpus. Our objective is to create a specialized corpus of the Bahraini Arabic dialect, which includes written texts as well as transcripts of audio files, belonging to a different genre (folktales, comedy shows, plays, cooking shows, etc.). The corpus comprises 620K words, carefully curated. We provide automatic morphological annotations of the full corpus using state-of-the-art morphosyntactic disambiguation for Gulf Arabic. We validate the quality of the annotations on a 7.6K word sample. We plan to make the annotated sample as well as the full corpus publicly available to support researchers interested in Arabic NLP.},
  url   = {https://aclanthology.org/2022.lrec-1.251}
}

@InProceedings{swanson-tyers:2022:LREC,
  author   = {Swanson, Daniel and Tyers, Francis},
  title   = {A Universal Dependencies Treebank of Ancient Hebrew},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2353--2361},
  abstract = {In this paper we present the initial construction of a Universal Dependencies treebank with morphological annotations of Ancient Hebrew containing portions of the Hebrew Scriptures (1579 sentences, 27K tokens) for use in comparative study with ancient translations and for analysis of the development of Hebrew syntax. We construct this treebank by applying a rule-based parser (300 rules) to an existing morphologically-annotated corpus with minimal constituency structure and manually verifying the output and present the results of this semi-automated annotation process and some of the annotation decisions made in the process of applying the UD guidelines to a new language.},
  url   = {https://aclanthology.org/2022.lrec-1.252}
}

@InProceedings{carvalho-EtAl:2022:LREC,
  author   = {Carvalho, Paula and Cunha, Bernardo and Santos, Raquel and Batista, Fernando and Ribeiro, Ricardo},
  title   = {Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2362--2370},
  abstract = {This paper introduces FIGHT, a dataset containing 63,450 tweets, posted before and after the official declaration of Covid-19 as a pandemic by online users in Portugal. This resource aims at contributing to the analysis of online hate speech targeting the most representative minorities in Portugal, namely the African descent and the Roma communities, and the LGBTQI community, the most commonly reported target of hate speech in social media at the European context. We present the methods for collecting the data, and provide insightful statistics on the distribution of tweets included in FIGHT, considering both the temporal and spatial dimensions. We also analyze the availability over time of tweets targeting the above-mentioned communities, distinguishing public, private and deleted tweets. We believe this study will contribute to better understand the dynamics of online hate speech in Portugal, particularly in adverse contexts, such as a pandemic outbreak, allowing the development of more informed and accurate hate speech resources for Portuguese.},
  url   = {https://aclanthology.org/2022.lrec-1.253}
}

@InProceedings{barkarson-steingrmsson-hafsteinsdttir:2022:LREC,
  author   = {Barkarson, Starkaður and Steingrímsson, Steinþór and Hafsteinsdóttir, Hildur},
  title   = {Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2371--2381},
  abstract = {The Icelandic Gigaword Corpus was first published in 2018. Since then new versions have been published annually, containing new texts from additional sources as well as from previous sources. This paper describes the evolution of the corpus in its first four years. All versions are made available under permissive licenses and with each new version the texts are annotated with the latest and most accurate tools. We show how the corpus has grown almost 50\% in size from the first version to the fourth and how it was restructured in order to better accommodate different meta-data for different subcorpora. Furthermore, other services have been set up to facilitate usage of the corpus for different use cases. These include a keyword-in-context concordance tool, an n-gram viewer, a word frequency database and pre-trained word embeddings.},
  url   = {https://aclanthology.org/2022.lrec-1.254}
}

@InProceedings{sileo-EtAl:2022:LREC,
  author   = {Sileo, Damien and Muller, Philippe and Van de Cruys, Tim and Pradel, Camille},
  title   = {A Pragmatics-Centered Evaluation Framework for Natural Language Understanding},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2382--2394},
  abstract = {New models for natural language understanding have recently made an unparalleled amount of progress, which has led some researchers to suggest that the models induce universal text representations. However, current benchmarks are predominantly targeting semantic phenomena; we make the case that pragmatics needs to take center stage in the evaluation of natural language understanding. We introduce PragmEval, a new benchmark for the evaluation of natural language understanding, that unites 11 pragmatics-focused evaluation datasets for English. PragmEval can be used as supplementary training data in a multi-task learning setup, and is publicly available, alongside the code for gathering and preprocessing the datasets. Using our evaluation suite, we show that natural language inference, a widely used pretraining task, does not result in genuinely universal representations, which presents a new challenge for multi-task learning.},
  url   = {https://aclanthology.org/2022.lrec-1.255}
}

@InProceedings{bothe-wermter:2022:LREC,
  author   = {Bothe, Chandrakant and Wermter, Stefan},
  title   = {Conversational Analysis of Daily Dialog Data using Polite Emotional Dialogue Acts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2395--2400},
  abstract = {Many socio-linguistic cues are used in conversational analysis, such as emotion, sentiment, and dialogue acts. One of the fundamental social cues is politeness, which linguistically possesses properties such as social manners useful in conversational analysis. This article presents findings of polite emotional dialogue act associations, where we can correlate the relationships between the socio-linguistic cues. We confirm our hypothesis that the utterances with the emotion classes Anger and Disgust are more likely to be impolite. At the same time, Happiness and Sadness are more likely to be polite. A less expectable phenomenon occurs with dialogue acts Inform and Commissive which contain more polite utterances than Question and Directive. Finally, we conclude on the

future work of these findings to extend the learning of social behaviours using politeness.},
  url   = {https://aclanthology.org/2022.lrec-1.256}
}

@InProceedings{chiarcos:2022:LREC,
  author   = {Chiarcos, Christian},
  title   = {Inducing Discourse Marker Inventories from Lexical Knowledge Graphs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2401--2412},
  abstract = {Discourse marker inventories are important tools for the development of both discourse parsers and corpora with discourse annotations. In this paper we explore the potential of massively multilingual lexical knowledge graphs to induce multilingual discourse marker lexicons using concept propagation methods as previously developed in the context of translation inference across dictionaries. Given one or multiple source languages with discourse marker inventories that discourse relations as senses of potential discourse markers, as well as a large number of bilingual dictionaries that link them -- directly or indirectly -- with the target language, we specifically study to what extent discourse marker induction can benefit from the integration of information from different sources, the impact of sense granularity and what limiting factors may need to be considered. Our study uses discourse marker inventories from nine European languages normalized against the discourse relation inventory of the Penn Discourse Treebank (PDTB), as well as three collections of machine-readable dictionaries with different characteristics, so that the interplay of a large number of factors can be studied.},
  url   = {https://aclanthology.org/2022.lrec-1.257}
}

@InProceedings{haghighatkhah-EtAl:2022:LREC,
  author   = {Haghighatkhah, Pantea and Fokkens, Antske and Sommerauer, Pia and Speckmann, Bettina and Verbeek, Kevin},
  title   = {Story Trees: Representing Documents using Topological Persistence},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2413--2429},
  abstract = {Topological Data Analysis (TDA) focuses on the inherent shape of (spatial) data. As such, it may provide useful methods to explore spatial representations of linguistic data (embeddings) which have become central in NLP. In this paper we aim

to introduce TDA to researchers in language technology. We use TDA to represent document structure as so-called story trees. Story trees are hierarchical representations created from semantic vector representations of sentences via persistent homology. They can be used to identify and clearly visualize prominent components of a story line. We showcase their potential by using story trees to create extractive summaries for news stories.},
  url   = {https://aclanthology.org/2022.lrec-1.258}
}

@InProceedings{zwittervitez-EtAl:2022:LREC,
  author   = {Zwitter Vitez, Ana and Brglez, Mojca and Robnik Šikonja, Marko and Škvorc, Tadej and Vezovnik, Andreja and Pollak, Senja},
  title   = {Extracting and Analysing Metaphors in Migration Media Discourse: towards a Metaphor Annotation Scheme},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2430--2439},
  abstract = {The study of metaphors in media discourse is an increasingly researched topic as media are an important shaper of social reality and metaphors are an indicator of how we think about certain issues through references to other things. We present a neural transfer learning method for detecting metaphorical sentences in Slovene and evaluate its performance on a gold standard corpus of metaphors (classification accuracy of 0.725), as well as on a sample of a domain specific corpus of migrations (precision of 0.40 for extracting domain metaphors and 0.74 if evaluated only on a set of migration related sentences). Based on empirical results and findings of our analysis, we propose a novel metaphor annotation scheme containing linguistic level, conceptual level, and stance information. The new scheme can be used for future metaphor annotations of other socially relevant topics.},
  url   = {https://aclanthology.org/2022.lrec-1.259}
}

@InProceedings{mohr-whrl-klinger:2022:LREC,
  author   = {Mohr, Isabelle and Wührl, Amelie and Klinger, Roman},
  title   = {CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {244--257},
  abstract = {During the first two years of the COVID-19 pandemic, large volumes of biomedical information concerning this new disease

have been published on social media. Some of this information can pose a real danger, particularly when false information is shared, for instance recommendations how to treat diseases without professional medical advice. Therefore, automatic fact-checking resources and systems developed specifically for medical domain are crucial. While existing fact-checking resources cover COVID-19 related information in news or quantify the amount of misinformation in tweets, there is no dataset providing fact-checked COVID-19 related Twitter posts with detailed annotations for biomedical entities, relations and relevant evidence. We contribute CoVERT, a fact-checked corpus of tweets with a focus on the domain of biomedicine and COVID-19 related (mis)information. The corpus consists of 300 tweets, each annotated with named entities and relations. We employ a novel crowdsourcing methodology to annotate all tweets with fact-checking labels and supporting evidence, which crowdworkers search for online. This methodology results in substantial inter-annotator agreement. Furthermore, we use the retrieved evidence extracts as part of a fact-checking pipeline, finding that the real-world evidence is more useful than the knowledge directly available in pretrained language models.},
  url   = {https://aclanthology.org/2022.lrec-1.26}
}

@InProceedings{flansmosemikkelsen-EtAl:2022:LREC,
  author   = {Flansmose Mikkelsen, Linea and Kinch, Oliver and
Jess Pedersen, Anders and Lacroix, Ophélie},
  title   = {DDisCo: A Discourse Coherence Dataset for Danish},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2440--2445},
  abstract = {To date, there has been no resource for studying
discourse coherence on real-world Danish texts. Discourse coherence
has mostly been approached with the assumption that incoherent texts
can be represented by coherent texts in which sentences have been
shuffled. However, incoherent real-world texts rarely resemble that.
We thus present DDisCo, a dataset including text from the Danish
Wikipedia and Reddit annotated for discourse coherence. We choose to
annotate real-world texts instead of relying on artificially
incoherent text for training and testing models. Then, we evaluate
the performance of several methods, including neural networks, on
the dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.260}
}

@InProceedings{mim-EtAl:2022:LREC,
  author   = {Mim, Farjana Sultana and Inoue, Naoya and Naito,
Shoichi and Singh, Keshav and Inui, Kentaro},
  title   = {LPAttack: A Feasible Annotation Scheme for Capturing
Logic Pattern of Attacks in Arguments},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2446--2459},
  abstract = {In argumentative discourse, persuasion is often achieved by refuting or attacking others' arguments. Attacking an argument is not always straightforward and often consists of complex rhetorical moves in which arguers may agree with a logic of an argument while attacking another logic. Furthermore, an arguer may neither deny nor agree with any logics of an argument, instead ignore them and attack the main stance of the argument by providing new logics and presupposing that the new logics have more value or importance than the logics presented in the attacked argument. However, there are no studies in computational argumentation that capture such complex rhetorical moves in attacks or the presuppositions or value judgments in them. To address this gap, we introduce LPAttack, a novel annotation scheme that captures the common modes and complex rhetorical moves in attacks along with the implicit presuppositions and value judgments. Our annotation study shows moderate inter-annotator agreement, indicating that human annotation for the proposed scheme is feasible. We publicly release our annotated corpus and the annotation guidelines.},
  url   = {https://aclanthology.org/2022.lrec-1.261}
}

@InProceedings{tracey-EtAl:2022:LREC2,
  author   = {Tracey, Jennifer and Rambow, Owen and Cardie, Claire and Dalton, Adam and Dang, Hoa Trang and Diab, Mona and Dorr, Bonnie and Guthrie, Louise and Markowska, Magdalena and Muresan, Smaranda and Prabhakaran, Vinodkumar and Shaikh, Samira and Strzalkowski, Tomek},
  title   = {BeSt: The Belief and Sentiment Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2460--2467},
  abstract = {We present the BeSt corpus, which records cognitive state: who believes what (i.e., factuality), and who has what sentiment towards what. This corpus is inspired by similar source-and-target corpora, specifically MPQA and FactBank. The corpus comprises two genres, newswire and discussion forums, in three languages, Chinese (Mandarin), English, and Spanish. The corpus is distributed through the LDC.},
  url   = {https://aclanthology.org/2022.lrec-1.262}
}

@InProceedings{wang-EtAl:2022:LREC1,
  author   = {Wang, Xintong and Schneider, Florian and Alacam, Özge and Chaudhury, Prateek and Biemann, Chris},

  title   = {MOTIF: Contextualized Images for Complex Words to Improve Human Reading},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2468--2477},
  abstract = {MOTIF (MultimOdal ConTextualized Images For Language Learners) is a multimodal dataset that consists of 1125 comprehension texts retrieved from Wikipedia Simple Corpus. Allowing multimodal processing or enriching the context with multimodal information has proven imperative for many learning tasks, specifically for second language (L2) learning. In this respect, several traditional NLP approaches can assist L2 readers in text comprehension processes, such as simplifying text or giving dictionary descriptions for complex words. As nicely stated in the well-known proverb, sometimes "a picture is worth a thousand words'' and an image can successfully complement the verbal message by enriching the representation, like in Pictionary books. This multimodal support can also assist on-the-fly text reading experience by providing a multimodal tool that chooses and displays the most relevant images for the difficult words, given the text context. This study mainly focuses on one of the key components to achieving this goal; collecting a multimodal dataset enriched with complex word annotation and validated image match.},
  url   = {https://aclanthology.org/2022.lrec-1.263}
}

@InProceedings{desisto-EtAl:2022:LREC,
  author   = {De Sisto, Mirella and Vandeghinste, Vincent and Egea Gómez, Santiago and De Coster, Mathieu and Shterionov, Dimitar and Saggion, Horacio},
  title   = {Challenges with Sign Language Datasets for Sign Language Recognition and Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2478--2487},
  abstract = {Sign Languages (SLs) are the primary means of communication for at least half a million people in Europe alone. However, the development of SL recognition and translation tools is slowed down by a series of obstacles concerning resource scarcity and standardization issues in the available data. The former challenge relates to the volume of data available for machine learning as well as the time required to collect and process new data. The latter obstacle is linked to the variety of the data, i.e., annotation formats are not unified and vary amongst different resources. The available data formats are often not suitable for machine learning, obstructing the provision of automatic tools based

on neural models. In the present paper, we give an overview of these challenges by comparing various SL corpora and SL machine learning datasets. Furthermore, we propose a framework to address the lack of standardization at format level, unify the available resources and facilitate SL research for different languages. Our framework takes ELAN files as inputs and returns textual and visual data ready to train SL recognition and translation models. We present a proof of concept, training neural translation models on the data produced by the proposed framework.},
  url   = {https://aclanthology.org/2022.lrec-1.264}
}

@InProceedings{mertz-EtAl:2022:LREC,
  author   = {Mertz, Clémence and BARREAUD, Vincent and Le Naour, Thibaut and Lolive, Damien and Gibet, Sylvie},
  title   = {A Low-Cost Motion Capture Corpus in French Sign Language for Interpreting Iconicity and Spatial Referencing Mechanisms},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2488--2497},
  abstract = {The automatic translation of sign language videos into transcribed texts is rarely approached in its whole, as it implies to finely model the grammatical mechanisms that govern these languages. The presented work is a first step towards the interpretation of French sign language (LSF) by specifically targeting iconicity and spatial referencing. This paper describes the LSF-SHELVES corpus as well as the original technology that was designed and implemented to collect it. Our goal is to use deep learning methods to circumvent the use of models in spatial referencing recognition. In order to obtain training material with sufficient variability, we designed a light-weight (and low-cost) capture protocol that enabled us to collect data from a large panel of LSF signers. This protocol involves the use of a portable device providing a 3D skeleton, and of a software developed specifically for this application to facilitate the post-processing of handshapes. The LSF-SHELVES includes simple and compound iconic and spatial dynamics, organized in 6 complexity levels, representing a total of 60 sequences signed by 15 LSF signers.},
  url   = {https://aclanthology.org/2022.lrec-1.265}
}

@InProceedings{verhagen-EtAl:2022:LREC,
  author   = {Verhagen, Marc and Lynch, Kelley and Rim, Kyeongmin and Pustejovsky, James},
  title   = {The CLAMS Platform at Work: Processing Audiovisual Data from the American Archive of Public Broadcasting},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2498--2506},
  abstract = {The Computational Linguistics Applications for Multimedia Services (CLAMS) platform provides access to computational content analysis tools for multimedia material. The version we present here is a robust update of an initial prototype implementation from 2019. The platform now sports a variety of image, video, audio and text processing tools that interact via a common multi-modal representation language named MMIF (Multi-Media Interchange Format). We describe the overall architecture, the MMIF format, some of the tools included in the platform, the process to set up and run a workflow, visualizations included in CLAMS, and evaluate aspects of the platform on data from the American Archive of Public Broadcasting, showing how CLAMS can add metadata to mass-digitized multimedia collections, metadata that are typically only available implicitly in now largely unsearchable digitized media in archives and libraries.},
  url   = {https://aclanthology.org/2022.lrec-1.266}
}

@InProceedings{reardon-EtAl:2022:LREC,
  author   = {Reardon, Carley and Paik, Sejin and Gao, Ge and Parekh, Meet and Zhao, Yanling and Guo, Lei and Betke, Margrit and Wijaya, Derry Tanti},
  title   = {BU-NEmo: an Affective Dataset of Gun Violence News},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2507--2516},
  abstract = {Given our society’s increased exposure to multimedia formats on social media platforms, efforts to understand how digital content impacts people’s emotions are burgeoning. As such, we introduce a U.S. gun violence news dataset that contains news headline and image pairings from 840 news articles with 15K high-quality, crowdsourced annotations on emotional responses to the news pairings. We created three experimental conditions for the annotation process: two with a single modality (headline or image only), and one multimodal (headline and image together). In contrast to prior works on affectively-annotated data, our dataset includes annotations on the dominant emotion experienced with the content, the intensity of the selected emotion and an open-ended, written component. By collecting annotations on different modalities of the same news content pairings, we explore the relationship between image and text influence on human emotional response. We offer initial analysis on our dataset, showing the nuanced affective differences that appear due to modality and individual factors such as political leaning and media consumption habits. Our dataset is made publicly available to facilitate future research in affective computing.},

  url   = {https://aclanthology.org/2022.lrec-1.267}
}

@InProceedings{reverdy-EtAl:2022:LREC,
  author   = {Reverdy, Justine and O'Connor Russell, Sam and Duquenne, Louise and Garaialde, Diego and Cowan, Benjamin R. and Harte, Naomi},
  title   = {RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2517--2527},
  abstract = {We present RoomReader, a corpus of multimodal, multiparty conversational interactions in which participants followed a collaborative student-tutor scenario designed to elicit spontaneous speech. The corpus was developed within the wider RoomReader Project to explore multimodal cues of conversational engagement and behavioural aspects of collaborative interaction in online environments. However, the corpus can be used to study a wide range of phenomena in online multimodal interaction. The publicly-shared corpus consists of over 8 hours of video and audio recordings from 118 participants in 30 gender-balanced sessions, in the “in-the-wild” online environment of Zoom. The recordings have been edited, synchronised, and fully transcribed. Student participants have been continuously annotated for engagement with a novel continuous scale. We provide questionnaires measuring engagement and group cohesion collected from the annotators, tutors and participants themselves. We also make a range of accompanying data available such as personality tests and behavioural assessments. The dataset and accompanying psychometrics present a rich resource enabling the exploration of a range of downstream tasks across diverse fields including linguistics and artificial intelligence. This could include the automatic detection of student engagement, analysis of group interaction and collaboration in online conversation, and the analysis of conversational behaviours in an online setting.},
  url   = {https://aclanthology.org/2022.lrec-1.268}
}

@InProceedings{sevilla-dazesteban-lahozbengoechea:2022:LREC,
  author   = {Sevilla, Antonio F. G. and Díaz Esteban, Alberto and Lahoz-Bengoechea, José María},
  title   = {Quevedo: Annotation and Processing of Graphical Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {2528--2535},
  abstract = {In this article, we present Quevedo, a software tool we have developed for the task of automatic processing of graphical languages. These are languages which use images to convey meaning, relying not only on the shape of symbols but also on their spatial arrangement in the page, and relative to each other. When presented in image form, these languages require specialized computational processing which is not the same as usually done either for natural language processing or for artificial vision. Quevedo enables this specialized processing, focusing on a data-based approach. As a command line application and library, it provides features for the collection and management of image datasets, and their machine learning recognition using neural networks and recognizer pipelines. This processing requires careful annotation of the source data, for which Quevedo offers an extensive and visual web-based annotation interface. In this article, we also briefly present a case study centered on the task of SignWriting recognition, the original motivation for writing the software. Quevedo is written in Python, and distributed freely under the Open Software License version 3.0.},
  url   = {https://aclanthology.org/2022.lrec-1.269}
}

@InProceedings{barbieri-espinosaanke-camachocollados:2022:LREC,
  author   = {Barbieri, Francesco and Espinosa Anke, Luis and Camacho-Collados, Jose},
  title   = {XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {258--266},
  abstract = {Language models are ubiquitous in current NLP, and their multilingual capacity has recently attracted considerable attention. However, current analyses have almost exclusively focused on (multilingual variants of) standard benchmarks, and have relied on clean pre-training and task-specific corpora as multilingual signals. In this paper, we introduce XLM-T, a model to train and evaluate multilingual language models in Twitter. In this paper we provide: (1) a new strong multilingual baseline consisting of an XLM-R (Conneau et al. 2020) model pre-trained on millions of tweets in over thirty languages, alongside starter code to subsequently fine-tune on a target task; and (2) a set of unified sentiment analysis Twitter datasets in eight different languages and a XLM-T model trained on this dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.27}
}

@InProceedings{saha-nayak-baumann:2022:LREC,
  author   = {Saha, Debjoy and Nayak, Shravan and Baumann, Timo},

  title   = {Merkel Podcast Corpus: A Multimodal Dataset Compiled from 16 Years of Angela Merkel’s Weekly Video Podcasts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2536--2540},
  abstract = {We introduce the Merkel Podcast Corpus, an audio-visual-text corpus in German collected from 16 years of (almost) weekly Internet podcasts of former German chancellor Angela Merkel. To the best of our knowledge, this is the first single speaker corpus in the German language consisting of audio, visual and text modalities of comparable size and temporal extent. We describe the methods used with which we have collected and edited the data which involves downloading the videos, transcripts and other metadata, forced alignment, performing active speaker recognition and face detection to finally curate the single speaker dataset consisting of utterances spoken by Angela Merkel. The proposed pipeline is general and can be used to curate other datasets of similar nature, such as talk show contents. Through various statistical analyses and applications of the dataset in talking face generation and TTS, we show the utility of the dataset. We argue that it is a valuable contribution to the research community, in particular, due to its realistic and challenging material at the boundary between prepared and spontaneous speech.},
  url   = {https://aclanthology.org/2022.lrec-1.270}
}

@InProceedings{mukushev-EtAl:2022:LREC,
  author   = {Mukushev, Medet and Kydyrbekova, Aigerim and Imashev, Alfarabi and Kimmelman, Vadim and Sandygulova, Anara},
  title   = {Crowdsourcing Kazakh-Russian Sign Language: FluentSigners-50},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2541--2547},
  abstract = {This paper presents the methodology we used to crowdsource a data collection of a new large-scale signer independent dataset for Kazakh-Russian Sign Language (KRSL) created for Sign Language Processing. By involving the Deaf community throughout the research process, we firstly designed a research protocol and then performed an efficient crowdsourcing campaign that resulted in a new FluentSigners-50 dataset. The FluentSigners-50 dataset consists of 173 sentences performed by 50 KRSL signers for 43,250 video samples. Dataset contributors recorded videos in real-life settings on various backgrounds using various devices such as smartphones and web cameras. Therefore, each dataset contribution has a varying distance to the camera, camera angles and aspect

ratio, video quality, and frame rates. Additionally, the proposed dataset contains a high degree of linguistic and inter-signer variability and thus is a better training set for recognizing a real-life signed speech. FluentSigners-50 is publicly available at https://krslproject.github.io/fluentsigners-50/},
  url   = {https://aclanthology.org/2022.lrec-1.271}
}

@InProceedings{nugues:2022:LREC,
  author   = {Nugues, Pierre},
  title   = {Connecting a French Dictionary from the Beginning of the 20th Century to Wikidata},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2548--2555},
  abstract = {The Petit Larousse illustré is a French dictionary first published in 1905. Its division in two main parts on language and on history and geography corresponds to a major milestone in French lexicography as well as a repository of general knowledge from this period. Although the value of many entries from 1905 remains intact, some descriptions now have a dimension that is more historical than contemporary. They are nonetheless significant to analyze and understand cultural representations from this time. A comparison with more recent information or a verification of these entries would require a tedious manual work. In this paper, we describe a new lexical resource, where we connected all the dictionary entries of the history and geography part to current data sources. For this, we linked each of these entries to a wikidata identifier. Using the wikidata links, we can automate more easily the identification, comparison, and verification of historically-situated representations. We give a few examples on how to process wikidata identifiers and we carried out a small analysis of the entities described in the dictionary to outline possible applications. The resource, i.e. the annotation of 20,245 dictionary entries with wikidata links, is available from GitHub (https://github.com/pnugues/petit_larousse_1905/)},
  url   = {https://aclanthology.org/2022.lrec-1.272}
}

@InProceedings{egg-kordoni:2022:LREC,
  author   = {Egg, Markus and Kordoni, Valia},
  title   = {Metaphor annotation for German},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2556--2562},
  abstract = {The paper presents current work on a German corpus

annotated for metaphor. Metaphors denote entities or situations that are in some sense similar to the literal referent, e.g., when "Handschrift" ‘signature’ is used in the sense of ‘distinguishing mark’ or the suppression of hopes is introduced by the verb "verschütten" ‘bury’. The corpus is part of a project on register, hence, includes material from different registers that represent register variation along a number of important dimensions, but we believe that it is of interest to research on metaphor in general. The corpus extends previous annotation initiatives in that it not only annotates the metaphoric expressions themselves but also their respective relevant contexts that trigger a metaphorical interpretation of the expressions. For the corpus, we developed extended annotation guidelines, which specifically focus not only on the identification of these metaphoric contexts but also analyse in detail specific linguistic challenges for metaphor annotation that emerge due to the grammar of German.},
  url   = {https://aclanthology.org/2022.lrec-1.273}
}

@InProceedings{kutuzov-EtAl:2022:LREC,
  author   = {Kutuzov, Andrey and Touileb, Samia and Mæhlum, Petter and Enstad, Tita and Wittemann, Alexandra},
  title   = {NorDiaChange: Diachronic Semantic Change Dataset for Norwegian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2563--2572},
  abstract = {We describe NorDiaChange: the first diachronic semantic change dataset for Norwegian. NorDiaChange comprises two novel subsets, covering about 80 Norwegian nouns manually annotated with graded semantic change over time. Both datasets follow the same annotation procedure and can be used interchangeably as train and test splits for each other. NorDiaChange covers the time periods related to pre- and post-war events, oil and gas discovery in Norway, and technological developments. The annotation was done using the DURel framework and two large historical Norwegian corpora. NorDiaChange is published in full under a permissive licence, complete with raw annotation data and inferred diachronic word usage graphs (DWUGs).},
  url   = {https://aclanthology.org/2022.lrec-1.274}
}

@InProceedings{gonalooliveira:2022:LREC,
  author   = {Gonçalo Oliveira, Hugo},
  title   = {Exploring Transformers for Ranking Portuguese Semantic Relations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2573--2582},
  abstract = {We explored transformer-based language models for ranking instances of Portuguese lexico-semantic relations. Weights were based on the likelihood of natural language sequences that transmitted the relation instances, and expectations were that they would be useful for filtering out noisier instances. However, after analysing the weights, no strong conclusions were taken. They are not correlated with redundancy, but are lower for instances with longer and more specific arguments, which may nevertheless be a consequence of their sensitivity to the frequency of such arguments. They did also not reveal to be useful when computing word similarity with network embeddings. Despite the negative results, we see the reported experiments and insights as another contribution for better understanding transformer language models like BERT and GPT, and we make the weighted instances publicly available for further research.},
  url   = {https://aclanthology.org/2022.lrec-1.275}
}

@InProceedings{ferret:2022:LREC,
  author   = {Ferret, Olivier},
  title   = {Building Static Embeddings from Contextual Ones: Is It Useful for Building Distributional Thesauri?},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2583--2590},
  abstract = {While contextual language models are now dominant in the field of Natural Language Processing, the representations they build at the token level are not always suitable for all uses. In this article, we propose a new method for building word or type-level embeddings from contextual models. This method combines the generalization and the aggregation of token representations. We evaluate it for a large set of English nouns from the perspective of the building of distributional thesauri for extracting semantic similarity relations. Moreover, we analyze the differences between static embeddings and type-level embeddings according to features such as the frequency of words or the type of semantic relations these embeddings account for, showing that the properties of these two types of embeddings can be complementary and exploited for further improving distributional thesauri.},
  url   = {https://aclanthology.org/2022.lrec-1.276}
}

@InProceedings{wang-EtAl:2022:LREC2,
  author   = {Wang, Yixiao and Bouraoui, Zied and Espinosa Anke, Luis and Schockaert, Steven},
  title   = {Sentence Selection Strategies for Distilling Word Embeddings from BERT},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2591--2600},
  abstract = {Many applications crucially rely on the availability of high-quality word vectors. To learn such representations, several strategies based on language models have been proposed in recent years. While effective, these methods typically rely on a large number of contextualised vectors for each word, which makes them impractical. In this paper, we investigate whether similar results can be obtained when only a few contextualised representations of each word can be used. To this end, we analyse a range of strategies for selecting the most informative sentences. Our results show that with a careful selection strategy, high-quality word vectors can be learned from as few as 5 to 10 sentences.},
  url   = {https://aclanthology.org/2022.lrec-1.277}
}

@InProceedings{baldissin-schlechtweg-schulteimwalde:2022:LREC,
  author   = {Baldissin, Gioia and Schlechtweg, Dominik and Schulte im Walde, Sabine},
  title   = {DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2601--2609},
  abstract = {We provide a novel dataset — DiaWUG — with judgements on diatopic lexical semantic variation for six Spanish variants in Europe and Latin America. In contrast to most previous meaning-based resources and studies on semantic diatopic variation, we collect annotations on semantic relatedness for Spanish target words in their contexts from both a semasiological perspective (i.e., exploring the meanings of a word given its form, thus including polysemy) and an onomasiological perspective (i.e., exploring identical meanings of words with different forms, thus including synonymy). In addition, our novel dataset exploits and extends the existing framework DURel for annotating word senses in context (Erk et al., 2013; Schlechtweg et al., 2018) and the framework-embedded Word Usage Graphs (WUGs) — which up to now have mainly be used for semasiological tasks and resources — in order to distinguish, visualize and interpret lexical semantic variation of contextualized words in Spanish from these two perspectives, i.e., semasiological and onomasiological language variation.},
  url   = {https://aclanthology.org/2022.lrec-1.278}
}

@InProceedings{chen-hulden:2022:LREC,

  author   = {Chen, Daniel and Hulden, Mans},
  title   = {My Case, For an Adposition: Lexical Polysemy of Adpositions and Case Markers in Finnish and Latin},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2610--2616},
  abstract = {Adpositions and case markers contain a high degree of polysemy and participate in unique semantic role configurations. We present a novel application of the SNACS supersense hierarchy to Finnish and Latin data by manually annotating adposition and case marker tokens in Finnish and Latin translations of Chapters IV-V of Le Petit Prince (The Little Prince). We evaluate the computational validity of the semantic role annotation categories by grouping raw, contextualized Multilingual BERT embeddings using k-means clustering.},
  url   = {https://aclanthology.org/2022.lrec-1.279}
}

@InProceedings{alhassan-zhang-schlegel:2022:LREC,
  author   = {Alhassan, Areej and Zhang, Jinkai and Schlegel, Viktor},
  title   = {‘Am I the Bad One’? Predicting the Moral Judgement of the Crowd Using Pre-trained Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {267--276},
  abstract = {Natural language processing (NLP) has been shown to perform well in various tasks, such as answering questions, ascertaining natural language inference and anomaly detection. However, there are few NLP-related studies that touch upon the moral context conveyed in text. This paper studies whether state-of-the-art, pre-trained language models are capable of passing moral judgments on posts retrieved from a popular Reddit user board. Reddit is a social discussion website and forum where posts are promoted by users through a voting system. In this work, we construct a dataset that can be used for moral judgement tasks by collecting data from the AITA? (Am I the A*******?) subreddit. To model our task, we harnessed the power of pre-trained language models, including BERT, RoBERTa, RoBERTa-large, ALBERT and Longformer. We then fine-tuned these models and evaluated their ability to predict the correct verdict as judged by users for each post in the datasets. RoBERTa showed relative improvements across the three datasets, exhibiting a rate of 87\% accuracy and a Matthews correlation coefficient (MCC) of 0.76, while the use of the Longformer model slightly improved the performance when used with longer sequences, achieving 87\% accuracy and 0.77 MCC.},

  url   = {https://aclanthology.org/2022.lrec-1.28}
}

@InProceedings{breit-revenko-blaschke:2022:LREC,
  author   = {Breit, Anna and Revenko, Artem and Blaschke, Narayani},
  title   = {WiC-TSV-de: German Word-in-Context Target-Sense-Verification Dataset and Cross-Lingual Transfer Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2617--2625},
  abstract = {Target Sense Verification (TSV) describes the binary disambiguation task of deciding whether the intended sense of a target word in a context corresponds to a given target sense. In this paper, we introduce WiC-TSV-de, a multi-domain dataset for German Target Sense Verification. While the training and development sets consist of domain-independent instances only, the test set contains domain-bound subsets, originating from four different domains, being Gastronomy, Medicine, Hunting, and Zoology. The domain-bound subsets incorporate adversarial examples such as in-domain ambiguous target senses and context-mixing (i.e., using the target sense in an out-of-domain context) which contribute to the challenging nature of the presented dataset. WiC-TSV-de allows for the development of sense-inventory-independent disambiguation models that can generalise their knowledge for different domain settings. By combining it with the original English WiC-TSV benchmark, we performed monolingual and cross-lingual analysis, where the evaluated baseline models were not able to solve the dataset to a satisfying degree, leaving a big gap to human performance.},
  url   = {https://aclanthology.org/2022.lrec-1.280}
}

@InProceedings{elboukkouri-EtAl:2022:LREC,
  author   = {El Boukkouri, Hicham and Ferret, Olivier and Lavergne, Thomas and Zweigenbaum, Pierre},
  title   = {Re-train or Train from Scratch? Comparing Pre-training Strategies of BERT in the Medical Domain},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2626--2633},
  abstract = {BERT models used in specialized domains all seem to be the result of a simple strategy: initializing with the original BERT and then resuming pre-training on a specialized corpus. This method yields rather good performance (e.g. BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019)). However, it seems reasonable to think that training directly

on a specialized corpus, using a specialized vocabulary, could result in more tailored embeddings and thus help performance. To test this hypothesis, we train BERT models from scratch using many configurations involving general and medical corpora. Based on evaluations using four different tasks, we find that the initial corpus only has a weak influence on the performance of BERT models when these are further pre-trained on a medical corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.281}
}

@InProceedings{orlando-EtAl:2022:LREC,
  author   = {Orlando, Riccardo and Conia, Simone and Faralli, Stefano and Navigli, Roberto},
  title   = {Universal Semantic Annotator: the First Unified API for WSD, SRL and Semantic Parsing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2634--2641},
  abstract = {In this paper, we present the Universal Semantic Annotator (USeA), which offers the first unified API for high-quality automatic annotations of texts in 100 languages through state-of-the-art systems for Word Sense Disambiguation, Semantic Role Labeling and Semantic Parsing. Together, such annotations can be used to provide users with rich and diverse semantic information, help second-language learners, and allow researchers to integrate explicit semantic knowledge into downstream tasks and real-world applications.},
  url   = {https://aclanthology.org/2022.lrec-1.282}
}

@InProceedings{wahle-EtAl:2022:LREC,
  author   = {Wahle, Jan Philip and Ruas, Terry and Mohammad, Saif and Gipp, Bela},
  title   = {D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2642--2651},
  abstract = {DBLP is the largest open-access repository of scientific articles on computer science and provides metadata associated with publications, authors, and venues. We retrieved more than 6 million publications from DBLP and extracted pertinent metadata (e.g., abstracts, author affiliations, citations) from the publication texts to create the DBLP Discovery Dataset (D3). D3 can be used to identify trends in research activity, productivity, focus, bias, accessibility, and impact of computer science research.

We present an initial analysis focused on the volume of computer science research (e.g., number of papers, authors, research activity), trends in topics of interest, and citation patterns. Our findings show that computer science is a growing research field (15\% annually), with an active and collaborative researcher community. While papers in recent years present more bibliographical entries in comparison to previous decades, the average number of citations has been declining. Investigating papers' abstracts reveals that recent topic trends are clearly reflected in D3. Finally, we list further applications of D3 and pose supplemental research questions. The D3 dataset, our findings, and source code are publicly available for research purposes.},
  url   = {https://aclanthology.org/2022.lrec-1.283}
}

@InProceedings{roussis-EtAl:2022:LREC2,
  author   = {Roussis, Dimitrios and Papavassiliou, Vassilis and Prokopidis, Prokopis and Piperidis, Stelios and Katsouros, Vassilis},
  title   = {SciPar: A Collection of Parallel Corpora from Scientific Abstracts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2652--2657},
  abstract = {This paper presents SciPar, a new collection of parallel corpora created from openly available metadata of bachelor theses, master theses and doctoral dissertations hosted in institutional repositories, digital libraries of universities and national archives. We describe first how we harvested and processed metadata from 86, mainly European, repositories to extract bilingual titles and abstracts, and then how we mined high quality sentence pairs in a wide range of scientific areas and sub-disciplines. In total, the resource includes 9.17 million segment alignments in 31 language pairs and is publicly available via the ELRC-SHARE repository. The bilingual corpora in this collection could prove valuable in various applications, such as cross-lingual plagiarism detection or adapting Machine Translation systems for the translation of scientific texts and academic writing in general, especially for language pairs which include English.},
  url   = {https://aclanthology.org/2022.lrec-1.284}
}

@InProceedings{gavidia-EtAl:2022:LREC,
  author   = {Gavidia, Martha and Lee, Patrick and Feldman, Anna and Peng, JIng},
  title   = {CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2658--2671},
  abstract = {Euphemisms have not received much attention in natural language processing, despite being an important element of polite and figurative language. Euphemisms prove to be a difficult topic, not only because they are subject to language change, but also because humans may not agree on what is a euphemism and what is not. Nonetheless, the first step to tackling the issue is to collect and analyze examples of euphemisms. We present a corpus of potentially euphemistic terms (PETs) along with example texts from the GloWbE corpus. Additionally, we present a subcorpus of texts where these PETs are not being used euphemistically, which may be useful for future applications. We also discuss the results of multiple analyses run on the corpus. Firstly, we find that sentiment analysis on the euphemistic texts supports that PETs generally decrease negative and offensive sentiment. Secondly, we observe cases of disagreement in an annotation task, where humans are asked to label PETs as euphemistic or not in a subset of our corpus text examples. We attribute the disagreement to a variety of potential reasons, including if the PET was a commonly accepted term (CAT).},
  url   = {https://aclanthology.org/2022.lrec-1.285}
}

@InProceedings{habash-EtAl:2022:LREC,
  author   = {Habash, Nizar and AbuOdeh, Muhammed and Taji, Dima and Faraj, Reem and El Gizuli, Jamila and Kallas, Omar},
  title   = {Camel Treebank: An Open Multi-genre Arabic Dependency Treebank},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2672--2681},
  abstract = {We present the Camel Treebank (CAMELTB), a 188K word open-source dependency treebank of Modern Standard and Classical Arabic. CAMELTB 1.0 includes 13 sub-corpora comprising selections of texts from pre-Islamic poetry to social media online commentaries, and covering a range of genres from religious and philosophical texts to news, novels, and student essays. The texts are all publicly available (out of copyright, creative commons, or under open licenses). The texts were morphologically tokenized and syntactically parsed automatically, and then manually corrected by a team of trained annotators. The annotations follow the guidelines of the Columbia Arabic Treebank (CATiB) dependency representation. We discuss our annotation process and guideline extensions, and we present some initial observations on lexical and syntactic differences among the annotated sub-corpora. This corpus will be publicly available to support and encourage research on Arabic NLP in general and on new, previously unexplored genres that are of interest to a wider spectrum of researchers, from historical

linguistics and digital humanities to computer-assisted language pedagogy.},
  url   = {https://aclanthology.org/2022.lrec-1.286}
}

@InProceedings{sotudeh-goharian-young:2022:LREC,
  author   = {Sotudeh, Sajad and Goharian, Nazli and Young, Zachary},
  title   = {MentSum: A Resource for Exploring Summarization of Mental Health Online Posts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2682--2692},
  abstract = {Mental health remains a significant challenge of public health worldwide. With increasing popularity of online platforms, many use the platforms to share their mental health conditions, express their feelings, and seek help from the community and counselors. Some of these platforms, such as Reachout, are dedicated forums where the users register to seek help. Others such as Reddit provide subreddits where the users publicly but anonymously post their mental health distress. Although posts are of varying length, it is beneficial to provide a short, but informative summary for fast processing by the counselors. To facilitate research in summarization of mental health online posts, we introduce Mental Health Summarization dataset, MentSum, containing over 24k carefully selected user posts from Reddit, along with their short user-written summary (called TLDR) in English from 43 mental health subreddits. This domain-specific dataset could be of interest not only for generating short summaries on Reddit, but also for generating summaries of posts on the dedicated mental health forums such as Reachout. We further evaluate both extractive and abstractive state-of-the-art summarization baselines in terms of Rouge scores, and finally conduct an in-depth human evaluation study of both user-written and system-generated summaries, highlighting challenges in this research.},
  url   = {https://aclanthology.org/2022.lrec-1.287}
}

@InProceedings{aumiller-gertz:2022:LREC,
  author   = {Aumiller, Dennis and Gertz, Michael},
  title   = {Klexikon: A German Dataset for Joint Summarization and Simplification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2693--2701},
  abstract = {Traditionally, Text Simplification is treated as a

monolingual translation task where sentences between source texts and their simplified counterparts are aligned for training. However, especially for longer input documents, summarizing the text (or dropping less relevant content altogether) plays an important role in the simplification process, which is currently not reflected in existing datasets. Simultaneously, resources for non-English languages are scarce in general and prohibitive for training new solutions. To tackle this problem, we pose core requirements for a system that can jointly summarize and simplify long source documents. We further describe the creation of a new dataset for joint Text Simplification and Summarization based on German Wikipedia and the German children's encyclopedia "Klexikon", consisting of almost 2,900 documents. We release a document-aligned version that particularly highlights the summarization aspect, and provide statistical evidence that this resource is well suited to simplification as well. Code and data are available on Github: https://github.com/dennlinger/klexikon},
  url   = {https://aclanthology.org/2022.lrec-1.288}
}

@InProceedings{hartl-kruschwitz:2022:LREC,
  author   = {Hartl, Philipp and Kruschwitz, Udo},
  title   = {Applying Automatic Text Summarization for Fake News Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2702--2713},
  abstract = {The distribution of fake news is not a new but a rapidly growing problem. The shift to news consumption via social media has been one of the drivers for the spread of misleading and deliberately wrong information, as in addition to its ease of use there is rarely any veracity monitoring. Due to the harmful effects of such fake news on society, the detection of these has become increasingly important. We present an approach to the problem that combines the power of transformer-based language models while simultaneously addressing one of their inherent problems. Our framework, CMTR-BERT, combines multiple text representations, with the goal of circumventing sequential limits and related loss of information the underlying transformer architecture typically suffers from. Additionally, it enables the incorporation of contextual information. Extensive experiments on two very different, publicly available datasets demonstrates that our approach is able to set new state-of-the-art performance benchmarks. Apart from the benefit of using automatic text summarization techniques we also find that the incorporation of contextual information contributes to performance gains.},
  url   = {https://aclanthology.org/2022.lrec-1.289}
}

@InProceedings{han-castroferreira-gardent:2022:LREC,

  author   = {Han, Kelvin and Castro Ferreira, Thiago and
Gardent, Claire},
  title   = {Generating Questions from Wikidata Triples},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {277--290},
  abstract = {Question generation from knowledge bases (or
knowledge base question generation, KBQG) is the task of generating
questions from structured database information, typically in the
form of triples representing facts. To handle rare entities and
generalize to unseen properties, previous work on KBQG resorted to
extensive, often ad-hoc pre- and post-processing of the input
triple. We revisit KBQG – using pre training, a new (triple,
question) dataset and taking question type into account – and show
that our approach outperforms previous work both in a standard and
in a zero-shot setting. We also show that the extended KBQG dataset
(also helpful for knowledge base question answering) we provide
allows not only for better coverage in terms of knowledge base (KB)
properties but also for increased output variability in that it
permits the generation of multiple questions from the same KB
triple.},
  url   = {https://aclanthology.org/2022.lrec-1.29}
}

@InProceedings{meisinger-trippel-zinn:2022:LREC,
  author   = {Meisinger, Nino and Trippel, Thorsten and Zinn,
Claus},
  title   = {Increasing CMDI's Semantic Interoperability with
schema.org},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2714--2720},
  abstract = {The CLARIN Concept Registry (CCR) is the common
semantic ground for most CMDI-based profiles to describe language-
related resources in the CLARIN universe. While the CCR supports
semantic interoperability within this universe, it does not extend
beyond it. The flexibility of CMDI, however, allows users to use
other term or concept registries when defining their metadata
components. In this paper, we describe our use of schema.org, a
light ontology used by many parties across disciplines.},
  url   = {https://aclanthology.org/2022.lrec-1.290}
}

@InProceedings{lange-aznar:2022:LREC,
  author   = {Lange, Herbert and Aznar, Jocelyn},
  title   = {RefCo and its Checker: Improving Language

Documentation Corpora's Reusability Through a Semi-Automatic Review Process},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2721--2729},
  abstract = {The QUEST (QUality ESTablished) project aims at ensuring the reusability of audio-visual datasets (Wamprechtshammer et al., 2022) by devising quality criteria and curating processes. RefCo (Reference Corpora) is an initiative within QUEST in collaboration with DoReCo (Documentation Reference Corpus, Paschen et al. (2020)) focusing on language documentation projects. Previously, Aznar and Seifart (2020) introduced a set of quality criteria dedicated to documenting fieldwork corpora. Based on these criteria, we establish a semi-automatic review process for existing and work-in-progress corpora, in particular for language documentation. The goal is to improve the quality of a corpus by increasing its reusability. A central part of this process is a template for machine-readable corpus documentation and automatic data verification based on this documentation. In addition to the documentation and automatic verification, the process involves a human review and potentially results in a RefCo certification of the corpus. For each of these steps, we provide guidelines and manuals. We describe the evaluation process in detail, highlight the current limits for automatic evaluation and how the manual review is organized accordingly.},
  url   = {https://aclanthology.org/2022.lrec-1.291}
}

@InProceedings{simon:2022:LREC,
  author   = {Simon, Gábor},
  title   = {Identification and Analysis of Personification in Hungarian: The PerSECorp project},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2730--2738},
  abstract = {Despite the recent findings on the conceptual and linguistic organization of personification, we have relatively little knowledge about its lexical patterns and grammatical templates. It is especially true in the case of Hungarian which has remained an understudied language regarding the constructions of figurative meaning generation. The present paper aims to provide a corpus-driven approach to personification analysis in the framework of cognitive linguistics. This approach is based on the building of a semi-automatically processed research corpus (the PerSE corpus) in which personifying linguistic structures are annotated manually. The present test version of the corpus consists of online car reviews

written in Hungarian (10468 words altogether): the texts were tokenized, lemmatized, morphologically analyzed, syntactically parsed, and PoS-tagged with the e-magyar NLP tool. For the identification of personifications, the adaptation of the MIPVU protocol was used and combined with additional analysis of semantic relations within personifying multi-word expressions. The paper demonstrates the structure of the corpus as well as the levels of the annotation. Furthermore, it gives an overview of possible data types emerging from the analysis: lexical pattern, grammatical characteristics, and the construction-like behavior of personifications in Hungarian.},
  url   = {https://aclanthology.org/2022.lrec-1.292}
}

@InProceedings{silvano-EtAl:2022:LREC,
  author   = {Silvano, Purificação and Damova, Mariana and Oleškevičienė, Giedrė Valūnaitė and Liebeskind, Chaya and Chiarcos, Christian and Trajanov, Dimitar and Truică, Ciprian-Octavian and Apostol, Elena-Simona and Baczkowska, Anna},
  title   = {ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2739--2749},
  abstract = {Discourse markers carry information about the discourse structure and organization, and also signal local dependencies or epistemological stance of speaker. They provide instructions on how to interpret the discourse, and their study is paramount to understand the mechanism underlying discourse organization. This paper presents a new language resource, an ISO-based annotated multilingual parallel corpus for discourse markers. The corpus comprises nine languages, Bulgarian, Lithuanian, German, European Portuguese, Hebrew, Romanian, Polish, and Macedonian, with English as a pivot language. In order to represent the meaning of the discourse markers, we propose an annotation scheme of discourse relations from ISO 24617-8 with a plug-in to ISO 24617-2 for communicative functions. We describe an experiment in which we applied the annotation scheme to assess its validity. The results reveal that, although some extensions are required to cover all the multilingual data, it provides a proper representation of discourse markers value. Additionally, we report some relevant contrastive phenomena concerning discourse markers interpretation and role in discourse. This first step will allow us to develop deep learning methods to identify and extract discourse relations and communicative functions, and to represent that information as Linguistic Linked Open Data (LLOD).},
  url   = {https://aclanthology.org/2022.lrec-1.293}
}

@InProceedings{gimenogmez-martnezhinarejos:2022:LREC,

  author   = {Gimeno-Gómez, David and Martínez-Hinarejos, Carlos-D.},
  title   = {LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2750--2758},
  abstract = {Speech is considered as a multi-modal process where hearing and vision are two fundamentals pillars. In fact, several studies have demonstrated that the robustness of Automatic Speech Recognition systems can be improved when audio and visual cues are combined to represent the nature of speech. In addition, Visual Speech Recognition, an open research problem whose purpose is to interpret speech by reading the lips of the speaker, has been a focus of interest in the last decades. Nevertheless, in order to estimate these systems in the currently Deep Learning era, large-scale databases are required. On the other hand, while most of these databases are dedicated to English, other languages lack sufficient resources. Thus, this paper presents a semi-automatically annotated audiovisual database to deal with unconstrained natural Spanish, providing 13 hours of data extracted from Spanish television. Furthermore, baseline results for both speaker-dependent and speaker-independent scenarios are reported using Hidden Markov Models, a traditional paradigm that has been widely used in the field of Speech Technologies.},
  url   = {https://aclanthology.org/2022.lrec-1.294}
}

@InProceedings{yun-kim-jung:2022:LREC,
  author   = {Yun, Hyeongu and Kim, Yongil and Jung, Kyomin},
  title   = {Modality Alignment between Deep Representations for Effective Video-and-Language Learning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2759--2770},
  abstract = {Video-and-Language learning, such as video question answering or video captioning, is the next challenge in the deep learning society, as it pursues the way how human intelligence perceives everyday life. These tasks require the ability of multi-modal reasoning which is to handle both visual information and text information simultaneously across time. In this point of view, a cross-modality attention module that fuses video representation and text representation takes a critical role in most recent approaches. However, existing Video-and-Language models merely compute the attention weights without considering the different characteristics of video modality and text modality. Such na¨ıve attention module

hinders the current models to fully enjoy the strength of cross-modality. In this paper, we propose a novel Modality Alignment method that benefits the cross-modality attention module by guiding it to easily amalgamate multiple modalities. Specifically, we exploit Centered Kernel Alignment (CKA) which was originally proposed to measure the similarity between two deep representations. Our method directly optimizes CKA to make an alignment between video and text embedding representations, hence it aids the cross-modality attention module to combine information over different modalities. Experiments on real-world Video QA tasks demonstrate that our method outperforms conventional multi-modal methods significantly with +3.57\% accuracy increment compared to the baseline in a popular benchmark dataset. Additionally, in a synthetic data environment, we show that learning the alignment with our method boosts the performance of the cross-modality attention.},
  url   = {https://aclanthology.org/2022.lrec-1.295}
}

@InProceedings{murat-koutsombogera-vogel:2022:LREC,
  author   = {Murat, Anais and Koutsombogera, Maria and Vogel, Carl},
  title   = {Mutual Gaze and Linguistic Repetition in a Multimodal Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2771--2780},
  abstract = {This paper investigates the correlation between mutual gaze and linguistic repetition, a form of alignment, which we take as evidence of mutual understanding. We focus on a multimodal corpus made of three-party conversations and explore the question of whether mutual gaze events correspond to moments of repetition or non-repetition. Our results, although mainly significant on word unigrams and bigrams, suggest positive correlations between the presence of mutual gaze and the repetitions of tokens, lemmas, or parts-of-speech, but negative correlations when it comes to paired levels of representation (tokens or lemmas associated with their part-of-speech). No compelling correlation is found with duration of mutual gaze. Results are strongest when ignoring punctuation as representations of pauses, intonation, etc. in counting aligned tokens.},
  url   = {https://aclanthology.org/2022.lrec-1.296}
}

@InProceedings{parisse-EtAl:2022:LREC,
  author   = {Parisse, Christophe and Blondel, Marion and Caët, Stéphanie and Danet, Claire and Vincent, Coralie and Morgenstern, Aliyah},
  title   = {Multidimensional Coding of Multimodal Languaging in Multi-Party Settings},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2781--2787},
  abstract = {In natural language settings, many interactions include more than two speakers, and real-life interpretation is based on all types of information available in all modalities. This constitutes a challenge for corpus-based analyses because the information in the audio and visual channels must be included in the coding. The goal of the DINLANG project is to tackle that challenge and analyze spontaneous interactions in family dinner settings (two adults and two to three children). The families use either French, or LSF (French sign language). Our aim is to compare how participants share language across the range of modalities found in vocal and visual languaging in coordination with dining. In order to pinpoint similarities and differences, we had to find a common coding tool for all situations (variations from one family to another) and modalities. Our coding procedure incorporates the use of the ELAN software. We created a template organized around participants, situations, and modalities, rather than around language forms. Spoken language transcription can be integrated, when it exists, but it is not mandatory. Data that has been created with another software can be injected in ELAN files if it is linked using time stamps. Analyses performed with the coded files rely on ELAN's structured search functionalities, which allow to achieve fine-grained temporal analyses and which can be completed by using spreadsheets or R language.},
  url   = {https://aclanthology.org/2022.lrec-1.297}
}

@InProceedings{kyjnek-EtAl:2022:LREC,
  author   = {Kyjánek, Lukáš and Lyashevskaya, Olga and Nedoluzhko, Anna and Vodolazsky, Daniil and Žabokrtský, Zdeněk},
  title   = {Constructing a Lexical Resource of Russian Derivational Morphology},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2788--2797},
  abstract = {Words of any language are to some extent related thought the ways they are formed. For instance, the verb 'exempl-ify' and the noun 'example-s' are both based on the word 'example', but the verb is derived from it, while the noun is inflected. In Natural Language Processing of Russian, the inflection is satisfactorily processed; however, there are only a few machine-trackable resources that capture derivations even though Russian has both of these morphological processes very rich. Therefore, we devote this paper to improving one of the methods of constructing such resources and to the application of the method to a Russian

lexicon, which results in the creation of the largest lexical resource of Russian derivational relations. The resulting database dubbed DeriNet.RU includes more than 300 thousand lexemes connected with more than 164 thousand binary derivational relations. To create such data, we combined the existing machine-learning methods that we improved to manage this goal. The whole approach is evaluated on our newly created data set of manual, parallel annotation. The resulting DeriNet.RU is freely available under an open license agreement.},
  url   = {https://aclanthology.org/2022.lrec-1.298}
}

@InProceedings{khishigsuren-EtAl:2022:LREC,
  author   = {Khishigsuren, Temuulen and Bella, Gábor and Batsuren, Khuyagbaatar and Freihat, Abed Alhakim and Chandran Nair, Nandu and Ganbold, Amarsanaa and Khalilia, Hadi and Chandrashekar, Yamini and giunchiglia, fausto},
  title   = {Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2798--2807},
  abstract = {This paper describes a method to enrich lexical resources with content relating to linguistic diversity, based on knowledge from the field of lexical typology. We capture the phenomenon of diversity through the notion of lexical gap and use a systematic method to infer gaps semi-automatically on a large scale, which we demonstrate on the kinship domain. The resulting free diversity-aware terminological resource consists of 198 concepts, 1,911 words, and 37,370 gaps in 699 languages. We see great potential in the use of resources such as ours for the improvement of a variety of cross-lingual NLP tasks, which we illustrate through an application in the evaluation of machine translation systems.},
  url   = {https://aclanthology.org/2022.lrec-1.299}
}

@InProceedings{park-EtAl:2022:LREC1,
  author   = {park, chanjun and Lee, Seolhwa and Seo, Jaehyung and Moon, Hyeonseok and Eo, Sugyeong and Lim, Heuiseok},
  title   = {Priming Ancient Korean Neural Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {22--28},
  abstract = {In recent years, there has been an increasing need for the restoration and translation of historical languages. In this study, we attempt to translate historical records in ancient Korean language based on neural machine translation (NMT). Inspired by

priming, a cognitive science theory that two different stimuli influence each other, we propose novel priming ancient-Korean NMT (AKNMT) using bilingual subword embedding initialization with structural property awareness in the ancient documents. Finally, we obtain state-of-the-art results in the AKNMT task. To the best of our knowledge, we confirm the possibility of developing a human-centric model that incorporates the concepts of cognitive science and analyzes the result from the perspective of interference and cognitive dissonance theory for the first time.},
  url   = {https://aclanthology.org/2022.lrec-1.3}
}

@InProceedings{muffo-cocco-bertino:2022:LREC,
  author   = {Muffo, Matteo and Cocco, Aldo and Bertino, Enrico},
  title   = {Evaluating Transformer Language Models on Arithmetic Operations Using Number Decomposition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {291--297},
  abstract = {In recent years, Large Language Models such as GPT-3 showed remarkable capabilities in performing NLP tasks in the zero and few shot settings. On the other hand, the experiments highlighted the difficulty of GPT-3 in carrying out tasks that require a certain degree of reasoning, such as arithmetic operations. In this paper we evaluate the ability of Transformer Language Models to perform arithmetic operations following a pipeline that, before performing computations, decomposes numbers in units, tens, and so on. We denote the models fine-tuned with this pipeline with the name Calculon and we test them in the task of performing additions, subtractions and multiplications on the same test sets of GPT-3. Results show an increase of accuracy of 63\% in the five-digit addition task. Moreover, we demonstrate the importance of the decomposition pipeline introduced, since fine-tuning the same Language Model without decomposing numbers results in 0\% accuracy in the five-digit addition task.},
  url   = {https://aclanthology.org/2022.lrec-1.30}
}

@InProceedings{paikens-EtAl:2022:LREC,
  author   = {Paikens, Peteris and Grasmanis, Mikus and Klints, Agute and Lokmane, Ilze and Pretkalniņa, Lauma and Rituma, Laura and Stāde, Madara and Strankale, Laine},
  title   = {Towards Latvian WordNet},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {2808--2815},
  abstract = {In this paper we describe our current work on creating a WordNet for Latvian based on the principles of the Princeton WordNet. The chosen methodology for word sense definition and sense linking is based on corpus evidence and the existing Tezaurs.lv online dictionary, ensuring a foundation that fits the Latvian language usage and existing linguistic tradition. We cover a wide set of semantic relations, including gradation sets. Currently the dataset consists of 6432 words linked in 5528 synsets, out of which 2717 synsets are considered fully completed as they have all the outgoing semantic links annotated, annotated with corpus examples for each sense and links to the English Princeton WordNet.},
  url   = {https://aclanthology.org/2022.lrec-1.300}
}

@InProceedings{liu-marco-gulla:2022:LREC,
  author   = {Liu, Peng and Marco, Cristina and Gulla, Jon Atle},
  title   = {Building Sentiment Lexicons for Mainland Scandinavian Languages Using Machine Translation and Sentence Embeddings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2816--2825},
  abstract = {This paper presents a simple but effective method to build sentiment lexicons for the three Mainland Scandinavian languages: Danish, Norwegian and Swedish. This method benefits from the English Sentiwordnet and a thesaurus in one of the target languages. Sentiment information from the English resource is mapped to the target languages by using machine translation and similarity measures based on sentence embeddings. A number of experiments with Scandinavian languages are performed in order to determine the best working sentence embedding algorithm for this task. A careful extrinsic evaluation on several datasets yields state-of-the-art results using a simple rule-based sentiment analysis algorithm. The resources are made freely available under an MIT License.},
  url   = {https://aclanthology.org/2022.lrec-1.301}
}

@InProceedings{nimb-EtAl:2022:LREC,
  author   = {Nimb, Sanni and Olsen, Sussi and Pedersen, Bolette and Troelsgård, Thomas},
  title   = {A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {2826--2832},
  abstract = {This paper describes how a newly published Danish sentiment lexicon with a high lexical coverage was compiled by use of lexicographic methods and based on the links between groups of words listed in semantic order in a thesaurus and the corresponding word sense descriptions in a comprehensive monolingual dictionary. The overall idea was to identify negative and positive sections in a thesaurus, extract the words from these sections and combine them with the dictionary information via the links. The annotation task of the dataset included several steps, and was based on the comparison of synonyms and near synonyms within a semantic field. In the cases where one of the words were included in the smaller Danish sentiment lexicon AFINN, its value there was used as inspiration and expanded to the synonyms when appropriate. In order to obtain a more practical lexicon with overall polarity values at lemma level, all the senses of the lemma were afterwards compared, taking into consideration dictionary information such as usage, style and frequency. The final lexicon contains 13,859 Danish polarity lemmas and includes morphological information. It is freely available at https://github.com/dsldk/danish-sentiment-lexicon (licence CC-BY-SA 4.0 International).},
  url   = {https://aclanthology.org/2022.lrec-1.302}
}

@InProceedings{chandrannair-EtAl:2022:LREC,
  author   = {Chandran Nair, Nandu and Velayuthan, Rajendran S. and Chandrashekar, Yamini and Bella, Gábor and giunchiglia, fausto},
  title   = {IndoUKC: A Concept-Centered Indian Multilingual Lexical Resource},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2833--2840},
  abstract = {We introduce the IndoUKC, a new multilingual lexical database comprised of eighteen Indian languages, with a focus on formally capturing words and word meanings specific to Indian languages and cultures. The IndoUKC reuses content from the existing IndoWordNet resource while providing a new model for the cross-lingual mapping of lexical meanings that allows for a richer, diversity-aware representation. Accordingly, beyond a thorough syntactic and semantic cleaning, the IndoWordNet lexical content has been thoroughly remodeled in order to allow a more precise expression of language-specific meaning. The resulting database is made available both for browsing through a graphical web interface and for download through the LiveLanguage data catalogue.},
  url   = {https://aclanthology.org/2022.lrec-1.303}
}

@InProceedings{kim-EtAl:2022:LREC2,
  author   = {Kim, Hyeondey and Kim, Seonhoon and KANG, INHO

and Kwak, Nojun and Fung, Pascale},
  title = {Korean Language Modeling via Syntactic Guide},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {2841--2849},
  abstract = {While pre-trained language models play a vital role in modern language processing tasks, but not every language can benefit from them. Most existing research on pre-trained language models focuses primarily on widely-used languages such as English, Chinese, and Indo-European languages. Additionally, such schemes usually require extensive computational resources alongside a large amount of data, which is infeasible for less-widely used languages. We aim to address this research niche by building a language model that understands the linguistic phenomena in the target language which can be trained with low-resources. In this paper, we discuss Korean language modeling, specifically methods for language representation and pre-training methods. With our Korean-specific language representation, we are able to build more powerful language models for Korean understanding, even with fewer resources. The paper proposes chunk-wise reconstruction of the Korean language based on a widely used transformer architecture and bidirectional language representation. We also introduce morphological features such as Part-of-Speech (PoS) into the language understanding by leveraging such information during the pre-training. Our experiment results prove that the proposed methods improve the model performance of the investigated Korean language understanding tasks.},
  url = {https://aclanthology.org/2022.lrec-1.304}
}

@InProceedings{zirikly-EtAl:2022:LREC,
  author = {Zirikly, Ayah and Desmet, Bart and Porcino, Julia and Camacho Maldonado, Jonathan and Ho, Pei-Shu and Jimenez Silva, Rafael and Sacco, Maryanne},
  title = {A Whole-Person Function Dictionary for the Mobility, Self-Care and Domestic Life Domains: a Seedset Expansion Approach},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {2850--2855},
  abstract = {Whole-person functional limitations in the areas of mobility, self-care and domestic life affect a majority of individuals with disabilities. Detecting, recording and monitoring such limitations would benefit those individuals, as well as research on whole-person functioning and general public health. Dictionaries of terms related to whole-person function would enable automated identification and extraction of relevant information.

However, no such terminologies currently exist, due in part to a lack of standardized coding and their availability mainly in free text clinical notes. In this paper, we introduce terminologies of whole-person function in the domains of mobility, self-care and domestic life, built and evaluated using a small set of manually annotated clinical notes, which provided a seedset that was expanded using a mix of lexical and deep learning approaches.},
  url   = {https://aclanthology.org/2022.lrec-1.305}
}

@InProceedings{giouli-EtAl:2022:LREC,
  author   = {Giouli, Voula and Vacalopoulou, Anna and Sidiropoulos, Nikolaos and Flouda, Christina and Doupas, Athanasios and Giannopoulos, Giorgos and Bikakis, Nikos and Kaffes, Vassilis and Stainhaouer, Gregory},
  title   = {Placing multi-modal, and multi-lingual Data in the Humanities Domain on the Map: the Mythotopia Geo-tagged Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2856--2864},
  abstract = {The paper gives an account of an infrastructure that will be integrated into a platform aimed at providing a multi-faceted experience to visitors of Northern Greece using mythology as a starting point. This infrastructure comprises a multi-lingual and multi-modal corpus (i.e., a corpus of textual data supplemented with images, and video) that belongs to the humanities domain along with a dedicated database (content management system) with advanced indexing, linking and search functionalities. We will present the corpus itself focusing on the content, the methodology adopted for its development, and the steps taken towards rendering it accessible via the database in a way that also facilitates useful visualizations. In this context, we tried to address three main challenges: (a) to add a novel annotation layer, namely geotagging, (b) to ensure the long-term maintenance of and accessibility to the highly heterogeneous primary data – even after the life cycle of the current project – by adopting a metadata schema that is compatible to existing standards; and (c) to render the corpus a useful resource to scholarly research in the digital humanities by adding a minimum set of linguistic annotations.},
  url   = {https://aclanthology.org/2022.lrec-1.306}
}

@InProceedings{ohya:2022:LREC,
  author   = {Ohya, Kazushi},
  title   = {An Architecture of resolving a multiple link path in a standoff-style data format to enhance the mobility of language resources},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2865--2873},
  abstract = {The present data formats proposed by authentic organizations are based on a so-called standoff-style data format in XML, which represents a semantic data model through an instance structure and a link structure. However, this type of data formats intended to enhance the power of representation of an XML format injures the mobility of data because an abstract data structure denoted by multiple link paths is hard to be converted into other data structures. This difficulty causes a problem in the reuse of data to convert into other data formats especially in a personal data management environment. In this paper, in order to compensate for the drawback, we propose a new concept of transforming a link structure to an instance structure on a new marked-up scheme. This approach to language data brings a new architecture of language data management to realize a personal data management environment in daily and long-life use.},
  url   = {https://aclanthology.org/2022.lrec-1.307}
}

@InProceedings{romberg-mark-escher:2022:LREC,
  author   = {Romberg, Julia and Mark, Laura and Escher, Tobias},
  title   = {A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2874--2883},
  abstract = {Political authorities in democratic countries regularly consult the public in order to allow citizens to voice their ideas and concerns on specific issues. When trying to evaluate the (often large number of) contributions by the public in order to inform decision-making, authorities regularly face challenges due to restricted resources. We identify several tasks whose automated support can help in the evaluation of public participation. These are i) the recognition of arguments, more precisely premises and their conclusions, ii) the assessment of the concreteness of arguments, iii) the detection of textual descriptions of locations in order to assign citizens' ideas to a spatial location, and iv) the thematic categorization of contributions. To enable future research efforts to develop techniques addressing these four tasks, we introduce the CIMT PartEval Corpus, a new publicly-available German-language corpus that includes several thousand citizen contributions from six mobility-related planning processes in five German municipalities. The corpus provides annotations for each of these tasks which have not been available in German for the domain of public participation before either at all or in this scope and

variety.},
  url   = {https://aclanthology.org/2022.lrec-1.308}
}

@InProceedings{lesage-EtAl:2022:LREC,
  author   = {Lesage, Jakob and Haynie, Hannah J. and Skirgård, Hedvig and Weber, Tobias and Witzlack-Makarevich, Alena},
  title   = {Overlooked Data in Typological Databases: What Grambank Teaches Us About Gaps in Grammars},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2884--2890},
  abstract = {Typological databases can contain a wealth of information beyond the collection of linguistic properties across languages. This paper shows how information often overlooked in typological databases can inform the research community about the state of description of the world’s languages. We illustrate this using Grambank, a morphosyntactic typological database covering 2,467 language varieties and based on 3,951 grammatical descriptions. We classify and quantify the comments that accompany coded values in Grambank. We then aggregate these comments and the coded values to derive a level of description for 17 grammatical domains that Grambank covers (negation, adnominal modification, participant marking, tense, aspect, etc.). We show that the description level of grammatical domains varies across space and time. Information about gaps and uncertainties in the descriptive knowledge of grammatical domains within and across languages is essential for a correct analysis of data in typological databases and for the study of grammatical diversity more generally. When collected in a database, such information feeds into disciplines that focus on primary data collection, such as grammaticography and language documentation.},
  url   = {https://aclanthology.org/2022.lrec-1.309}
}

@InProceedings{naraki-sakai-hayashi:2022:LREC,
  author   = {Naraki, Yuji and Sakai, Tetsuya and Hayashi, Yoshihiko},
  title   = {Evaluating the Effects of Embedding with Speaker Identity Information in Dialogue Summarization},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {298--304},
  abstract = {Automatic dialogue summarization is a task used to succinctly summarize a dialogue transcript while correctly linking the speakers and their speech, which distinguishes this task from a

conventional document summarization. To address this issue and reduce the ``who said what''-related errors in a summary, we propose embedding the speaker identity information in the input embedding into the dialogue transcript encoder. Unlike the speaker embedding proposed by Gu et al. (2020), our proposal takes into account the informativeness of position embedding. By experimentally comparing several embedding methods, we confirmed that the scores of ROUGE and a human evaluation of the generated summaries were substantially increased by embedding speaker information at the less informative part of the fixed position embedding with sinusoidal functions.},
  url   = {https://aclanthology.org/2022.lrec-1.31}
}

@InProceedings{mccarthy-dore:2022:LREC,
  author   = {McCarthy, Arya D. and Dore, Giovanna Maria Dora},
  title   = {Hong Kong: Longitudinal and Synchronic Characterisations of Protest News between 1998 and 2020},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2891--2900},
  abstract = {This paper showcases the utility and timeliness of the Hong Kong Protest News Dataset, a highly curated collection of news articles from diverse news sources, to investigate longitudinal and synchronic news characterisations of protests in Hong Kong between 1998 and 2020. The properties of the dataset enable us to apply natural language processing to its 4522 articles and thereby study patterns of journalistic practice across newspapers. This paper sheds light on whether depth and/or manner of reporting changed over time, and if so, in what ways, or in response to what. In its focus and methodology, this paper helps bridge the gap between “validity-focused methodological debates” and the use of computational methods of analysis in the social sciences.},
  url   = {https://aclanthology.org/2022.lrec-1.310}
}

@InProceedings{volk-EtAl:2022:LREC,
  author   = {Volk, Martin and Fischer, Lukas and Scheurer, Patricia and Schroffenegger, Bernard Silvan and Schwitter, Raphael and Ströbel, Phillip and Suter, Benjamin},
  title   = {Nunc profana tractemus. Detecting Code-Switching in a Large Corpus of 16th Century Letters},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2901--2908},
  abstract = {This paper is based on a collection of 16th century letters from and to the Zurich reformer Heinrich Bullinger. Around

12,000 letters of this exchange have been preserved, out of which 3100 have been professionally edited, and another 5500 are available as provisional transcriptions. We have investigated code-switching in these 8600 letters, first on the sentence-level and then on the word-level. In this paper we give an overview of the corpus and its language mix (mostly Early New High German and Latin, but also French, Greek, Italian and Hebrew). We report on our experiences with a popular language identifier and present our results when training an alternative identifier on a very small training corpus of only 150 sentences per language. We use the automatically labeled sentences in order to bootstrap a word-based language classifier which works with high accuracy. Our research around the corpus building and annotation involves automatic handwritten text recognition, text normalisation for ENH German, and machine translation from medieval Latin into modern German.},
  url   = {https://aclanthology.org/2022.lrec-1.311}
}

@InProceedings{mikulov-EtAl:2022:LREC,
  author   = {Mikulová, Marie and Straka, Milan and Štěpánek, Jan and Štěpánková, Barbora and Hajic, Jan},
  title   = {Quality and Efficiency of Manual Annotation: Pre-annotation Bias},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2909--2918},
  abstract = {This paper presents an analysis of annotation using an automatic pre-annotation for a mid-level annotation complexity task - dependency syntax annotation. It compares the annotation efforts made by annotators using a pre-annotated version (with a high-accuracy parser) and those made by fully manual annotation. The aim of the experiment is to judge the final annotation quality when pre-annotation is used. In addition, it evaluates the effect of automatic linguistically-based (rule-formulated) checks and another annotation on the same data available to the annotators, and their influence on annotation quality and efficiency. The experiment confirmed that the pre-annotation is an efficient tool for faster manual syntactic annotation which increases the consistency of the resulting annotation without reducing its quality.},
  url   = {https://aclanthology.org/2022.lrec-1.312}
}

@InProceedings{ocal-EtAl:2022:LREC2,
  author   = {Ocal, Mustafa and Radas, Antonela and Hummer, Jared and Megerdoomian, Karine and Finlayson, Mark},
  title   = {A Comprehensive Evaluation and Correction of the TimeBank Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2919--2927},
  abstract = {TimeML is an annotation scheme for capturing temporal information in text. The developers of TimeML built the TimeBank corpus to both validate the scheme and provide a rich dataset of events, temporal expressions, and temporal relationships for training and testing temporal analysis systems. In our own work we have been developing methods aimed at TimeML graphs for detecting (and eventually automatically correcting) temporal inconsistencies, extracting timelines, and assessing temporal indeterminacy. In the course of this investigation we identified numerous previously unrecognized issues in the TimeBank corpus, including multiple violations of TimeML annotation guide rules, incorrectly disconnected temporal graphs, as well as inconsistent, redundant, missing, or otherwise incorrect annotations. We describe our methods for detecting and correcting these problems, which include: (a) automatic guideline checking (109 violations); (b) automatic inconsistency checking (65 inconsistent files); (c) automatic disconnectivity checking (625 incorrect breakpoints); and (d) manual comparison with the output of state-of-the-art automatic annotators to identify missing annotations (317 events, 52 temporal expressions). We provide our code as well as a set of patch files that can be applied to the TimeBank corpus to produce a corrected version for use by other researchers in the field.},
  url   = {https://aclanthology.org/2022.lrec-1.313}
}

@InProceedings{tripodi-blloshmi-levissullam:2022:LREC,
  author   = {Tripodi, Rocco and Blloshmi, Rexhina and Levis Sullam, Simon},
  title   = {Evaluating Multilingual Sentence Representation Models in a Real Case Scenario},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2928--2939},
  abstract = {In this paper, we present an evaluation of sentence representation models on the paraphrase detection task. The evaluation is designed to simulate a real-world problem of plagiarism and is based on one of the most important cases of forgery in modern history: the so-called "Protocols of the Elders of Zion''. The sentence pairs for the evaluation are taken from the infamous forged text "Protocols of the Elders of Zion'' (Protocols) by unknown authors; and by "Dialogue in Hell between Machiavelli and Montesquieu'' by Maurice Joly. Scholars have demonstrated that the first text plagiarizes from the second, indicating all the forged parts on qualitative grounds. Following this evidence, we organized the rephrased texts and asked native speakers to quantify the level of similarity between each pair. We used this material to evaluate

sentence representation models in two languages: English and French, and on three tasks: similarity correlation, paraphrase identification, and paraphrase retrieval. Our evaluation aims at encouraging the development of benchmarks based on real-world problems, as a means to prevent problems connected to AI hypes, and to use NLP technologies for social good. Through our evaluation, we are able to confirm that the infamous Protocols are actually a plagiarized text but, as we will show, we encounter several problems connected with the convoluted nature of the task, that is very different from the one reported in standard benchmarks of paraphrase detection and sentence similarity. Code and data available at https://github.com/roccotrip/protocols.},
  url   = {https://aclanthology.org/2022.lrec-1.314}
}

@InProceedings{baledent-EtAl:2022:LREC,
  author   = {Baledent, Anaëlle and Mathet, Yann and Widlöcher, Antoine and Couronne, Christophe and Manguin, Jean-Luc},
  title   = {Validity, Agreement, Consensuality and Annotated Data Quality},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2940--2948},
  abstract = {Reference annotated (or gold-standard) datasets are required for various common tasks such as training for machine learning systems or system validation. They are necessary to analyse or compare occurrences or items annotated by experts, or to compare objects resulting from any computational process to objects annotated by experts. But, even if reference annotated gold-standard corpora are required, their production is known as a difficult problem, from both a theoretical and practical point of view. Many studies devoted to theses issues conclude that multi-annotation is most of the time a necessity. That inter-annotator agreement measure, which is required to check the reliability of data and the reproducibility of an annotation task, and thus to establish a gold standard, is another thorny problem. Fine analysis of available metrics for this specific task then becomes essential. Our work is part of this effort and more precisely focuses on several problems, which are rarely discussed, although they are intrinsically linked with the interpretation of metrics. In particular, we focus here on the complex relations between agreement and reference (of which agreement among annotators is supposed to be an indicator), and the emergence of consensus. We also introduce the notion of consensuality as another relevant indicator.},
  url   = {https://aclanthology.org/2022.lrec-1.315}
}

@InProceedings{mdhaffar-EtAl:2022:LREC,
  author   = {mdhaffar, salima and Pelloin, Valentin and Caubrière, Antoine and Laperriere, Gaëlle and Ghannay, Sahar

and Jabaian, Bassam and Camelin, Nathalie and Estève, Yannick},
  title   = {Impact Analysis of the Use of Speech and Language Models Pretrained by Self-Supersivion for Spoken Language Understanding},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2949--2956},
  abstract = {Pretrained models through self-supervised learning have been recently introduced for both acoustic and language modeling. Applied to spoken language understanding tasks, these models have shown their great potential by improving the state-of-the-art performances on challenging benchmark datasets. In this paper, we present an error analysis reached by the use of such models on the French MEDIA benchmark dataset, known as being one of the most challenging benchmarks for the slot filling task among all the benchmarks accessible to the entire research community. One year ago, the state-of-art system reached a Concept Error Rate (CER) of 13.6\% through the use of a end-to-end neural architecture. Some months later, a cascade approach based on the sequential use of a fine-tuned wav2vec2.0 model and a fine-tuned BERT model reaches a CER of 11.2\%. This significant improvement raises questions about the type of errors that remain difficult to treat, but also about those that have been corrected using these models pre-trained through self-supervision learning on a large amount of data. This study brings some answers in order to better understand the limits of such models and open new perspectives to continue improving the performance.},
  url   = {https://aclanthology.org/2022.lrec-1.316}
}

@InProceedings{kurihara-kawahara-shibata:2022:LREC,
  author   = {Kurihara, Kentaro and Kawahara, Daisuke and Shibata, Tomohide},
  title   = {JGLUE: Japanese General Language Understanding Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2957--2966},
  abstract = {To develop high-performance natural language understanding (NLU) models, it is necessary to have a benchmark to evaluate and analyze NLU ability from various perspectives. While the English NLU benchmark, GLUE, has been the forerunner, benchmarks are now being released for languages other than English, such as CLUE for Chinese and FLUE for French; but there is no such benchmark for Japanese. We build a Japanese NLU benchmark, JGLUE, from scratch without translation to measure the general NLU ability in Japanese.

We hope that JGLUE will facilitate NLU research in Japanese.},
  url   = {https://aclanthology.org/2022.lrec-1.317}
}

@InProceedings{akhlaghi-EtAl:2022:LREC,
  author   = {Akhlaghi, Elham and Auðunardóttir, Ingibjörg Iða and Bączkowska, Anna and Bédi, Branislav and Beedar, Hakeem and Berthelsen, Harald and Chua, Cathy and Cucchiarin, Catia and Habibi, Hanieh and Horváthová, Ivana and Ikeda, Junta and Maizonniaux, Christèle and Ní Chiaráin, Neasa and Raheb, Chadi and Rayner, Manny and Sloan, John and Tsourakis, Nikos and Yao, Chunlin},
  title   = {Using the LARA Little Prince to compare human and TTS audio quality},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2967--2975},
  abstract = {A popular idea in Computer Assisted Language Learning (CALL) is to use multimodal annotated texts, with annotations typically including embedded audio and translations, to support L2 learning through reading. An important question is how to create good quality audio, which can be done either through human recording or by a Text-To-Speech (TTS) engine. We may reasonably expect TTS to be quicker and easier, but human to be of higher quality. Here, we report a study using the open source LARA platform and ten languages. Samples of audio totalling about five minutes, representing the same four passages taken from LARA versions of Saint-Exupèry's "Le petit prince", were provided for each language in both human and TTS form; the passages were chosen to instantiate the 2x2 cross product of the conditions {dialogue, not-dialogue} and {humour, not-humour}. 251 subjects used a web form to compare human and TTS versions of each item and rate the voices as a whole. For the three languages where TTS did best, English, French and Irish, the evidence from this study and the previous one it extended suggest that TTS audio is now pedagogically adequate and roughly comparable with a non-professional human voice in terms of exemplifying correct pronunciation and prosody. It was however still judged substantially less natural and less pleasant to listen to. No clear evidence was found to support the hypothesis that dialogue and humour pose special problems for TTS. All data and software will be made freely available.},
  url   = {https://aclanthology.org/2022.lrec-1.318}
}

@InProceedings{emmery-EtAl:2022:LREC,
  author   = {Emmery, Chris and Kádár, Ákos and Chrupała, Grzegorz and Daelemans, Walter},
  title   = {Cyberbullying Classifiers are Sensitive to Model-Agnostic Perturbations},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2976--2988},
  abstract = {A limited amount of studies investigates the role of model-agnostic adversarial behavior in toxic content classification. As toxicity classifiers predominantly rely on lexical cues, (deliberately) creative and evolving language-use can be detrimental to the utility of current corpora and state-of-the-art models when they are deployed for content moderation. The less training data is available, the more vulnerable models might become. This study is, to our knowledge, the first to investigate the effect of adversarial behavior and augmentation for cyberbullying detection. We demonstrate that model-agnostic lexical substitutions significantly hurt classifier performance. Moreover, when these perturbed samples are used for augmentation, we show models become robust against word-level perturbations at a slight trade-off in overall task performance. Augmentations proposed in prior work on toxicity prove to be less effective. Our results underline the need for such evaluations in online harm areas with small corpora.},
  url   = {https://aclanthology.org/2022.lrec-1.319}
}

@InProceedings{monsen-rennes:2022:LREC,
  author   = {Monsen, Julius and Rennes, Evelina},
  title   = {Perceived Text Quality and Readability in Extractive and Abstractive Summaries},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {305--312},
  abstract = {We present results from a study investigating how users perceive text quality and readability in extractive and abstractive summaries. We trained two summarisation models on Swedish news data and used these to produce summaries of articles. With the produced summaries, we conducted an online survey in which the extractive summaries were compared to the abstractive summaries in terms of fluency, adequacy and simplicity. We found statistically significant differences in perceived fluency and adequacy between abstractive and extractive summaries but no statistically significant difference in simplicity. Extractive summaries were preferred in most cases, possibly due to the types of errors the summaries tend to have.},
  url   = {https://aclanthology.org/2022.lrec-1.32}
}

@InProceedings{ellison-same:2022:LREC,
  author   = {Ellison, T. Mark and Same, Fahime},
  title   = {Constructing Distributions of Variation in Referring

Expression Type from Corpora for Model Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2989--2997},
  abstract = {The generation of referring expressions (REs) is a non-deterministic task. However, the algorithms for the generation of REs are standardly evaluated against corpora of written texts which include only one RE per each reference. Our goal in this work is firstly to reproduce one of the few studies taking the distributional nature of the RE generation into account. We add to this work, by introducing a method for exploring variation in human RE choice on the basis of longitudinal corpora - substantial corpora with a single human judgement (in the process of composition) per RE. We focus on the prediction of RE types, proper name, description and pronoun. We compare evaluations made against distributions over these types with evaluations made against parallel human judgements. Our results show agreement in the evaluation of learning algorithms against distributions constructed from parallel human evaluations and from longitudinal data.},
  url   = {https://aclanthology.org/2022.lrec-1.320}
}

@InProceedings{perevalov-EtAl:2022:LREC,
  author   = {Perevalov, Aleksandr and Yan, Xi and Kovriguina, Liubov and Jiang, Longquan and Both, Andreas and Usbeck, Ricardo},
  title   = {Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {2998--3007},
  abstract = {Data-driven systems need to be evaluated to establish trust in the scientific approach and its applicability. In particular, this is true for Knowledge Graph (KG) Question Answering (QA), where complex data structures are made accessible via natural-language interfaces. Evaluating the capabilities of these systems has been a driver for the community for more than ten years while establishing different KGQA benchmark datasets. However, comparing different approaches is cumbersome. The lack of existing and curated leaderboards leads to a missing global view over the research field and could inject mistrust into the results. In particular, the latest and most-used datasets in the KGQA community, LC-QuAD and QALD, miss providing central and up-to-date points of trust. In this paper, we survey and analyze a wide range of evaluation results with significant coverage of 100 publications and 98 systems from the last decade. We provide a new central and open leaderboard for any

KGQA benchmark dataset as a focal point for the community - https://kgqa.github.io/leaderboard/. Our analysis highlights existing problems during the evaluation of KGQA systems. Thus, we will point to possible improvements for future evaluations.},
  url   = {https://aclanthology.org/2022.lrec-1.321}
}

@InProceedings{takase-okazaki:2022:LREC,
  author   = {Takase, Sho and Okazaki, Naoaki},
  title   = {Multi-Task Learning for Cross-Lingual Abstractive Summarization},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3008--3016},
  abstract = {We present a multi-task learning framework for cross-lingual abstractive summarization to augment training data. Recent studies constructed pseudo cross-lingual abstractive summarization data to train their neural encoder-decoders. Meanwhile, we introduce existing genuine data such as translation pairs and monolingual abstractive summarization data into training. Our proposed method, Transum, attaches a special token to the beginning of the input sentence to indicate the target task. The special token enables us to incorporate the genuine data into the training data easily. The experimental results show that Transum achieves better performance than the model trained with only pseudo cross-lingual summarization data. In addition, we achieve the top ROUGE score on Chinese-English and Arabic-English abstractive summarization. Moreover, Transum also has a positive effect on machine translation. Experimental results indicate that Transum improves the performance from the strong baseline, Transformer, in Chinese-English, Arabic-English, and English-Japanese translation datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.322}
}

@InProceedings{castilho:2022:LREC,
  author   = {Castilho, Sheila},
  title   = {How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3017--3025},
  abstract = {This paper analyses how much context span is necessary to solve different context-related issues, namely, reference, ellipsis, gender, number, lexical ambiguity, and terminology when translating from English into Portuguese. We use the DELA corpus, which consists of 60 documents and six different

domains (subtitles, literary, news, reviews, medical, and legislation). We find that the shortest context span to disambiguate issues can appear in different positions in the document including preceding, following, global, world knowledge. Moreover, the average length depends on the issue types as well as the domain. Moreover, we show that the standard approach of relying on only two preceding sentences as context might not be enough depending on the domain and issue types.},
  url   = {https://aclanthology.org/2022.lrec-1.323}
}

@InProceedings{gete-EtAl:2022:LREC,
  author   = {Gete, Harritxu and Etchegoyhen, Thierry and Ponce, David and Labaka, Gorka and Aranberri, Nora and Corral, Ander and Saralegi, Xabier and Ellakuria, Igor and Martin, Maite},
  title   = {TANDO: A Corpus for Document-level Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3026--3037},
  abstract = {Document-level Neural Machine Translation aims to increase the quality of neural translation models by taking into account contextual information. Properly modelling information beyond the sentence level can result in improved machine translation output in terms of coherence, cohesion and consistency. Suitable corpora for context-level modelling are necessary to both train and evaluate context-aware systems, but are still relatively scarce. In this work we describe TANDO, a document-level corpus for the under-resourced Basque-Spanish language pair, which we share with the scientific community. The corpus is composed of parallel data from three different domains and has been prepared with context-level information. Additionally, the corpus includes contrastive test sets for fine-grained evaluations of gender and register contextual phenomena on both source and target language sides. To establish the usefulness of the corpus, we trained and evaluated baseline Transformer models and context-aware variants based on context concatenation. Our results indicate that the corpus is suitable for fine-grained evaluation of document-level machine translation systems.},
  url   = {https://aclanthology.org/2022.lrec-1.324}
}

@InProceedings{degibertbonet-EtAl:2022:LREC1,
  author   = {de Gibert Bonet, Ona and Goenaga, Iakes and Armengol-Estapé, Jordi and Perez-de-Viñaspre, Olatz and Parra Escartín, Carla and Sanchez, Marina and Pinnis, Mārcis and Labaka, Gorka and Melero, Maite},
  title   = {Unsupervised Machine Translation in Real-World Scenarios},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3038--3047},
  abstract = {In this work, we present the work that has been carried on in the MT4All CEF project and the resources that it has generated by leveraging recent research carried out in the field of unsupervised learning. In the course of the project 18 monolingual corpora for specific domains and languages have been collected, and 12 bilingual dictionaries and translation models have been generated. As part of the research, the unsupervised MT methodology based only on monolingual corpora (Artetxe et al., 2017) has been tested on a variety of languages and domains. Results show that in specialised domains, when there is enough monolingual in-domain data, unsupervised results are comparable to those of general domain supervised translation, and that, at any rate, unsupervised techniques can be used to boost results whenever very little data is available.},
  url   = {https://aclanthology.org/2022.lrec-1.325}
}

@InProceedings{ashida-kim-seunghun:2022:LREC,
  author   = {Ashida, Mana and Kim, Jin-Dong and Seunghun, Lee},
  title   = {COVID-19 Mythbusters in World Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3048--3055},
  abstract = {This paper introduces a multi-lingual database containing translated texts of COVID-19 mythbusters. The database has translations into 115 languages as well as the original English texts, of which the original texts are published by World Health Organization (WHO). This paper then presents preliminary analyses on latin-alphabet-based texts to see the potential of the database as a resource for multilingual linguistic analyses. The analyses on latin-alphabet-based texts gave interesting insights into the resource. While the amount of translated texts in each language was small, character bi-grams with normalization (lowercasing and removal of diacritics) was turned out to be an effective proxy for measuring the similarity of the languages, and the affinity ranking of language pairs could be obtained. Additionally, the hierarchical clustering analysis is performed using the character bigram overlap ratio of every possible pair of languages. The result shows the cluster of Germanic languages, Romance languages, and Southern Bantu languages. In sum, the multilingual database not only offers fixed set of materials in numerous languages, but also serves as a preliminary tool to identify the language family using text-based

similarity measure of bigram overlap ratio.},
  url   = {https://aclanthology.org/2022.lrec-1.326}
}

@InProceedings{armengolestap-degibertbonet-melero:2022:LREC,
  author   = {Armengol-Estapé, Jordi and de Gibert Bonet, Ona and Melero, Maite},
  title   = {On the Multilingual Capabilities of Very Large-Scale English Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3056--3068},
  abstract = {Generative Pre-trained Transformers (GPTs) have recently been scaled to unprecedented sizes in the history of machine learning. These models, solely trained on the language modeling objective, have been shown to exhibit outstanding zero, one, and few-shot learning capabilities in a number of different tasks. Nevertheless, aside from anecdotal experiences, little is known regarding their multilingual capabilities, given the fact that the pre-training corpus is almost entirely composed of English text. In this work, we investigate its potential and limits in three tasks: extractive question-answering, text summarization and natural language generation for five different languages, as well as the effect of scale in terms of model size. Our results show that GPT-3 can be almost as useful for many languages as it is for English, with room for improvement if optimization of the tokenization is addressed.},
  url   = {https://aclanthology.org/2022.lrec-1.327}
}

@InProceedings{karakanta-EtAl:2022:LREC,
  author   = {Karakanta, Alina and Buet, François and Cettolo, Mauro and Yvon, François},
  title   = {Evaluating Subtitle Segmentation for End-to-end Generation Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3069--3078},
  abstract = {Subtitles appear on screen as short pieces of text, segmented based on formal constraints (length) and syntactic/semantic criteria. Subtitle segmentation can be evaluated with sequence segmentation metrics against a human reference. However, standard segmentation metrics cannot be applied when systems generate outputs different than the reference, e.g. with end-to-end subtitling systems. In this paper, we study ways to conduct reference-based evaluations of segmentation accuracy irrespective of

the textual content. We first conduct a systematic analysis of existing metrics for evaluating subtitle segmentation. We then introduce Sigma, a Subtitle Segmentation Score derived from an approximate upper-bound of BLEU on segmentation boundaries, which allows us to disentangle the effect of good segmentation from text quality. To compare Sigma with existing metrics, we further propose a boundary projection method from imperfect hypotheses to the true reference. Results show that all metrics are able to reward high quality output but for similar outputs system ranking depends on each metric's sensitivity to error type. Our thorough analyses suggest Sigma is a promising segmentation candidate but its reliability over other segmentation metrics remains to be validated through correlations with human judgements.},
  url   = {https://aclanthology.org/2022.lrec-1.328}
}

@InProceedings{rapp:2022:LREC,
  author   = {Rapp, Reinhard},
  title   = {Using Semantic Role Labeling to Improve Neural Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3079--3083},
  abstract = {Despite impressive progress in machine translation in recent years, it has occasionally been argued that current systems are still mainly based on pattern recognition and that further progress may be possible by using text understanding techniques, thereby e.g. looking at semantics of the type “Who is doing what to whom?”. In the current research we aim to take a small step into this direction. Assuming that semantic role labeling (SRL) grasps some of the relevant semantics, we automatically annotate the source language side of a standard parallel corpus, namely Europarl, with semantic roles. We then train a neural machine translation (NMT) system using the annotated corpus on the source language side, and the original unannotated corpus on the target language side. New text to be translated is first annotated by the same SRL system and then fed into the translation system. We compare the results to those of a baseline NMT system trained with unannotated text on both sides and find that the SRL-based system yields small improvements in terms of BLEU scores for each of the four language pairs under investigation, involving English, French, German, Greek and Spanish.},
  url   = {https://aclanthology.org/2022.lrec-1.329}
}

@InProceedings{mei-EtAl:2022:LREC,
  author   = {Mei, Alex and Kabir, Anisha and Bapat, Rukmini and Judge, John and Sun, Tony and Wang, William Yang},
  title   = {Learning to Prioritize: Precision-Driven Sentence Filtering for Long Text Summarization},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {313--318},
  abstract = {Neural text summarization has shown great potential in recent years. However, current state-of-the-art summarization models are limited by their maximum input length, posing a challenge to summarizing longer texts comprehensively. As part of a layered summarization architecture, we introduce PureText, a simple yet effective pre-processing layer that removes low- quality sentences in articles to improve existing summarization models. When evaluated on popular datasets like WikiHow and Reddit TIFU, we show up to 3.84 and 8.57 point ROUGE-1 absolute improvement on the full test set and the long article subset, respectively, for state-of-the-art summarization models such as BertSum and BART. Our approach provides downstream models with higher-quality sentences for summarization, improving overall model performance, especially on long text articles.},
  url   = {https://aclanthology.org/2022.lrec-1.33}
}

@InProceedings{bandyopadhyay-EtAl:2022:LREC,
  author   = {Bandyopadhyay, Dibyanayan and De, Arkadipta and Gain, Baban and Saikh, Tanik and Ekbal, Asif},
  title   = {A Deep Transfer Learning Method for Cross-Lingual Natural Language Inference},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3084--3092},
  abstract = {Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), has been one of the central tasks in Artificial Intelligence (AI) and Natural Language Processing (NLP). RTE between the two pieces of texts is a crucial problem, and it adds further challenges when involving two different languages, i.e., in the cross-lingual scenario. This paper proposes an effective transfer learning approach for cross-lingual NLI. We perform experiments on English-Hindi language pairs in the cross-lingual setting to find out that our novel loss formulation could enhance the performance of the baseline model by up to 2\%. To assess the effectiveness of our method further, we perform additional experiments on every possible language pair using four European languages, namely French, German, Bulgarian, and Turkish, on top of XNLI dataset. Evaluation results yield up to 10\% performance improvement over the respective baseline models, in some cases surpassing the state-of-the-art (SOTA). It is also to be noted that our proposed model has 110M parameters which is much lesser than the SOTA model having 220M parameters. Finally, we argue that

our transfer learning-based loss objective is model agnostic and thus can be used with other deep learning-based architectures for cross-lingual NLI.},
  url   = {https://aclanthology.org/2022.lrec-1.330}
}

@InProceedings{shardlow-alvamanchego:2022:LREC,
  author   = {Shardlow, Matthew and Alva-Manchego, Fernando},
  title   = {Simple TICO-19: A Dataset for Joint Translation and Simplification of COVID-19 Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3093--3102},
  abstract = {Specialist high-quality information is typically first available in English, and it is written in a language that may be difficult to understand by most readers. While Machine Translation technologies contribute to mitigate the first issue, the translated content will most likely still contain complex language. In order to investigate and address both problems simultaneously, we introduce Simple TICO-19, a new language resource containing manual simplifications of the English and Spanish portions of the TICO-19 corpus for Machine Translation of COVID-19 literature. We provide an in-depth description of the annotation process, which entailed designing an annotation manual and employing four annotators (two native English speakers and two native Spanish speakers) who simplified over 6,000 sentences from the English and Spanish portions of the TICO-19 corpus. We report several statistics on the new dataset, focusing on analysing the improvements in readability from the original texts to their simplified versions. In addition, we propose baseline methodologies for automatically generating the simplifications, translations and joint translation and simplifications contained in our dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.331}
}

@InProceedings{adjali-morin-zweigenbaum:2022:LREC,
  author   = {Adjali, Omar and Morin, Emmanuel and Zweigenbaum, Pierre},
  title   = {Building Comparable Corpora for Assessing Multi-Word Term Alignment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3103--3112},
  abstract = {Recent work has demonstrated the importance of dealing with Multi-Word Terms (MWTs) in several Natural Language Processing applications. In particular, MWTs pose serious challenges

for alignment and machine translation systems because of their syntactic and semantic properties. Thus, developing algorithms that handle MWTs is becoming essential for many NLP tasks. However, the availability of bilingual and more generally multi-lingual resources is limited, especially for low-resourced languages and in specialized domains. In this paper, we propose an approach for building comparable corpora and bilingual term dictionaries that help evaluate bilingual term alignment in comparable corpora. To that aim, we exploit parallel corpora to perform automatic bilingual MWT extraction and comparable corpus construction. Parallel information helps to align bilingual MWTs and makes it easier to build comparable specialized sub-corpora. Experimental validation on an existing dataset and on manually annotated data shows the interest of the proposed methodology.},
  url   = {https://aclanthology.org/2022.lrec-1.332}
}

@InProceedings{slmundsdttir-EtAl:2022:LREC,
  author   = {Sólmundsdóttir, Agnes and Guðmundsdóttir, Dagbjört and Stefánsdóttir, Lilja Björk and Ingason, Anton},
  title   = {Mean Machine Translations: On Gender Bias in Icelandic Machine Translations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3113--3121},
  abstract = {This paper examines machine bias in language technology. Machine bias can affect machine learning algorithms when language models trained on large corpora include biased human decisions or reflect historical or social inequities, e.g. regarding gender and race. The focus of the paper is on gender bias in machine translation and we discuss a study conducted on Icelandic translations in the translation systems Google Translate and Vélþýðing.is. The results show a pattern which corresponds to certain societal ideas about gender. For example it seems to depend on the meaning of adjectives referring to people whether they appear in the masculine or feminine form. Adjectives describing positive personality traits were more likely to appear in masculine gender whereas the negative ones frequently appear in feminine gender. However, the opposite applied to appearance related adjectives. These findings unequivocally demonstrate the importance of being vigilant towards technology so as not to maintain societal inequalities and outdated views – especially in today’s digital world.},
  url   = {https://aclanthology.org/2022.lrec-1.333}
}

@InProceedings{enayet-sukthankar:2022:LREC,
  author   = {Enayet, Ayesha and Sukthankar, Gita},
  title   = {An Analysis of Dialogue Act Sequence Similarity Across Multiple Domains},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3122--3130},
  abstract = {This paper presents an analysis of how dialogue act sequences vary across different datasets in order to anticipate the potential degradation in the performance of learned models during domain adaptation. We hypothesize the following: 1) dialogue sequences from related domains will exhibit similar n-gram frequency distributions 2) this similarity can be expressed by measuring the average Hamming distance between subsequences drawn from different datasets. Our experiments confirm that when dialogue acts sequences from two datasets are dissimilar they lie further away in embedding space, making it possible to train a classifier to discriminate between them even when the datasets are corrupted with noise. We present results from eight different datasets: SwDA, AMI (DialSum), GitHub, Hate Speech, Teams, Diplomacy Betrayal, SAMsum, and Military (Army). Our datasets were collected from many types of human communication including strategic planning, informal discussion, and social media exchanges. Our methodology provides intuition on the generalizability of dialogue models trained on different datasets. Based on our analysis, it is problematic to assume that machine learning models trained on one type of discourse will generalize well to other settings, due to contextual differences.},
  url   = {https://aclanthology.org/2022.lrec-1.334}
}

@InProceedings{okahisa-EtAl:2022:LREC,
  author   = {Okahisa, Taro and Tanaka, Ribeka and Kodama, Takashi and Huang, Yin Jou and Kurohashi, Sadao},
  title   = {Constructing a Culinary Interview Dialogue Corpus with Video Conferencing Tool},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3131--3139},
  abstract = {Interview is an efficient way to elicit knowledge from experts of different domains. In this paper, we introduce CIDC, an interview dialogue corpus in the culinary domain in which interviewers play an active role to elicit culinary knowledge from the cooking expert. The corpus consists of 308 interview dialogues (each about 13 minutes in length), which add up to a total of 69,000 utterances. We use a video conferencing tool for data collection, which allows us to obtain the facial expressions of the interlocutors as well as the screen-sharing contents. To understand the impact of the interlocutors' skill level, we divide the experts into "semi-professionals'" and "enthusiasts" and the interviewers into "skilled interviewers" and "unskilled interviewers." For

quantitative analysis, we report the statistics and the results of the post-interview questionnaire. We also conduct qualitative analysis on the collected interview dialogues and summarize the salient patterns of how interviewers elicit knowledge from the experts. The corpus serves the purpose to facilitate future research on the knowledge elicitation mechanism in interview dialogues.},
  url   = {https://aclanthology.org/2022.lrec-1.335}
}

@InProceedings{yusupujiang-ginzburg:2022:LREC,
  author   = {Yusupujiang, Zulipiye and Ginzburg, Jonathan},
  title   = {UgChDial: A Uyghur Chat-based Dialogue Corpus for Response Space Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3140--3149},
  abstract = {In this paper, we introduce a carefully designed and collected language resource: UgChDial -- a Uyghur dialogue corpus based on a chatroom environment. The Uyghur Chat-based Dialogue Corpus (UgChDial) is divided into two parts: (1). Two-party dialogues and (2). Multi-party dialogues. We ran a series of 25, 120-minutes each, two-party chat sessions, totaling 7323 turns and 1581 question-response pairs. We created 16 different scenarios and topics to gather these two-party conversations. The multi-party conversations were compiled from chitchats in general channels as well as free chats in topic-oriented public channels, yielding 5588 unique turns and 838 question-response pairs. The initial purpose of this corpus is to study query-response pairs in Uyghur, building on an existing fine-grained response space taxonomy for English. We provide here initial annotation results on the Uyghur response space classification task using UgChDial.},
  url   = {https://aclanthology.org/2022.lrec-1.336}
}

@InProceedings{sudo-EtAl:2022:LREC,
  author   = {Sudo, Saki and Asano, Kyoshiro and Mitsuda, Koh and Higashinaka, Ryuichiro and Takeuchi, Yugo},
  title   = {A Speculative and Tentative Common Ground Handling for Efficient Composition of Uncertain Dialogue},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3150--3157},
  abstract = {This study investigates how the grounding process is composed and explores new interaction approaches that adapt to human cognitive processes that have not yet been significantly studied. The results of an experiment indicate that grounding through

dialogue is mutually accepted among participants through holistic expressions and suggest that common ground among participants may not necessarily be formed in a bottom-up way through analytic expressions. These findings raise the possibility of a promising new approach to creating a human-like dialogue system that may be more suitable for natural human communication.},
  url   = {https://aclanthology.org/2022.lrec-1.337}
}

@InProceedings{aguirre-EtAl:2022:LREC,
  author   = {Aguirre, Maia and García-Sardiña, Laura and Serras, Manex and Méndez, Ariane and López, Jacobo},
  title   = {BaSCo: An Annotated Basque-Spanish Code-Switching Corpus for Natural Language Understanding},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3158--3163},
  abstract = {The main objective of this work is the elaboration and public release of BaSCo, the first corpus with annotated linguistic resources encompassing Basque-Spanish code-switching. The mixture of Basque and Spanish languages within the same utterance is popularly referred to as Euskañol, a widespread phenomenon among bilingual speakers in the Basque Country. Thus, this corpus has been created to meet the demand of annotated linguistic resources in Euskañol in research areas such as multilingual dialogue systems. The presented resource is the result of translating to Euskañol a compilation of texts in Basque and Spanish that were used for training the Natural Language Understanding (NLU) models of several task-oriented bilingual chatbots. Those chatbots were meant to answer specific questions associated with the administration, fiscal, and transport domains. In addition, they had the transverse potential to answer to greetings, requests for help, and chit-chat questions asked to chatbots. BaSCo is a compendium of 1377 tagged utterances with every sample annotated at three levels: (i) NLU semantic labels, considering intents and entities, (ii) code-switching proportion, and (iii) domain of origin.},
  url   = {https://aclanthology.org/2022.lrec-1.338}
}

@InProceedings{kraus-wagner-minker:2022:LREC,
  author   = {Kraus, Matthias and Wagner, Nicolas and Minker, Wolfgang},
  title   = {ProDial -- An Annotated Proactive Dialogue Act Corpus for Conversational Assistants using Crowdsourcing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {3164--3173},
  abstract = {Robots will eventually enter our daily lives and assist with a variety of tasks. Especially in the household domain, robots may become indispensable helpers by overtaking tedious tasks, e.g. keeping the place tidy. Their effectiveness and efficiency, however, depend on their ability to adapt to our needs, routines, and personal characteristics. Otherwise, they may not be accepted and trusted in our private domain. For enabling adaptation, the interaction between a human and a robot needs to be personalized. Therefore, the robot needs to collect personal information from the user. However, it is unclear how such sensitive data can be collected in an understandable way without losing a user's trust in the system. In this paper, we present a conversational approach for explicitly collecting personal user information using natural dialogue. For creating a sound interactive personalization, we have developed an empathy-augmented dialogue strategy. In an online study, the empathy-augmented strategy was compared to a baseline dialogue strategy for interactive personalization. We have found the empathy-augmented strategy to perform notably friendlier. Overall, using dialogue for interactive personalization has generally shown positive user reception.},
  url   = {https://aclanthology.org/2022.lrec-1.339}
}

@InProceedings{ishigaki-EtAl:2022:LREC,
  author   = {Ishigaki, Tatsuya and Nishino, Suzuko and Washino, Sohei and Igarashi, Hiroki and Nagai, Yukari and Washida, Yuichi and Murai, Akihiko},
  title   = {Automating Horizon Scanning in Future Studies},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {319--327},
  abstract = {We introduce document retrieval and comment generation tasks for automating horizon scanning. This is an important task in the field of futurology that collects sufficient information for predicting drastic societal changes in the mid- or long-term future. The steps used are: 1) retrieving news articles that imply drastic changes, and 2) writing subjective comments on each article for others' ease of understanding. As a first step in automating these tasks, we create a dataset that contains 2,266 manually collected news articles with comments written by experts. We analyze the collected documents and comments regarding characteristic words, the distance to general articles, and contents in the comments. Furthermore, we compare several methods for automating horizon scanning. Our experiments show that 1) manually collected articles are different from general articles regarding the words used and semantic distances, 2) the contents in the comment can be classified into several categories, and 3) a supervised model trained on our dataset achieves a better performance. The contributions are: 1) we propose document retrieval and comment

generation tasks for horizon scanning, 2) create and analyze a new dataset, and 3) report the performance of several models and show that comment generation tasks are challenging.},
  url   = {https://aclanthology.org/2022.lrec-1.34}
}

@InProceedings{nedoluzhko-EtAl:2022:LREC1,
  author   = {Nedoluzhko, Anna and Singh, Muskaan and Hledíková, Marie and Ghosal, Tirthankar and Bojar, Ondřej},
  title   = {ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3174--3182},
  abstract = {Taking minutes is an essential component of every meeting, although the goals, style, and procedure of this activity (``minuting'' for short) can vary. Minuting is a rather unstructured writing activity and is affected by who is taking the minutes and for whom the intended minutes are. With the rise of online meetings, automatic minuting would be an important benefit for the meeting participants as well as for those who might have missed the meeting. However, automatically generating meeting minutes is a challenging problem due to a variety of factors including the quality of automatic speech recorders (ASRs), availability of public meeting data, subjective knowledge of the minuter, etc. In this work, we present the first of its kind dataset on \textit{Automatic Minuting}. We develop a dataset of English and Czech technical project meetings which consists of transcripts generated from ASRs, manually corrected, and minuted by several annotators. Our dataset, AutoMin, consists of 113 (English) and 53 (Czech) meetings, covering more than 160 hours of meeting content. Upon acceptance, we will publicly release (aaa.bbb.ccc) the dataset as a set of meeting transcripts and minutes, excluding the recordings for privacy reasons. A unique feature of our dataset is that most meetings are equipped with more than one minute, each created independently. Our corpus thus allows studying differences in what people find important while taking the minutes. We also provide baseline experiments for the community to explore this novel problem further. To the best of our knowledge \textbf{AutoMin} is probably the first resource on minuting in English and also in a language other than English (Czech).}},
  url   = {https://aclanthology.org/2022.lrec-1.340}
}

@InProceedings{fraser-kiritchenko-nejadgholi:2022:LREC,
  author   = {Fraser, Kathleen C. and Kiritchenko, Svetlana and Nejadgholi, Isar},
  title   = {Extracting Age-Related Stereotypes from Social Media Texts},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3183--3194},
  abstract = {Age-related stereotypes are pervasive in our society, and yet have been under-studied in the NLP community. Here, we present a method for extracting age-related stereotypes from Twitter data, generating a corpus of 300,000 over-generalizations about four contemporary generations (baby boomers, generation X, millennials, and generation Z), as well as "old" and "young" people more generally. By employing word-association metrics, semi-supervised topic modelling, and density-based clustering, we uncover many common stereotypes as reported in the media and in the psychological literature, as well as some more novel findings. We also observe trends consistent with the existing literature, namely that definitions of "young" and "old" age appear to be context-dependent, stereotypes for different generations vary across different topics (e.g., work versus family life), and some age-based stereotypes are distinct from generational stereotypes. The method easily extends to other social group labels, and therefore can be used in future work to study stereotypes of different social categories. By better understanding how stereotypes are formed and spread, and by tracking emerging stereotypes, we hope to eventually develop mitigating measures against such biased statements.},
  url   = {https://aclanthology.org/2022.lrec-1.341}
}

@InProceedings{alvarezmellado-lignos:2022:LREC,
  author   = {Alvarez-Mellado, Elena and Lignos, Constantine},
  title   = {Borrowing or Codeswitching? Annotating for Finer-Grained Distinctions in Language Mixing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3195--3201},
  abstract = {We present a new corpus of Twitter data annotated for codeswitching and borrowing between Spanish and English. The corpus contains 9,500 tweets annotated at the token level with codeswitches, borrowings, and named entities. This corpus differs from prior corpora of codeswitching in that we attempt to clearly define and annotate the boundary between codeswitching and borrowing and do not treat common "internet-speak" (lol, etc.) as codeswitching when used in an otherwise monolingual context. The result is a corpus that enables the study and modeling of Spanish-English borrowing and codeswitching on Twitter in one dataset. We present baseline scores for modeling the labels of this corpus using Transformer-based language models. The annotation itself is released with a CC BY 4.0 license, while the text it applies to is distributed in compliance with the Twitter terms of service.},

  url   = {https://aclanthology.org/2022.lrec-1.342}
}

@InProceedings{uban-chulvi-rosso:2022:LREC,
  author   = {Uban, Ana Sabina and Chulvi, Berta and Rosso,
Paolo},
  title   = {Multi-Aspect Transfer Learning for Detecting Low
Resource Mental Disorders on Social Media},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3202--3219},
  abstract = {Mental disorders are a serious and increasingly
relevant public health issue. NLP methods have the potential to
assist with automatic mental health disorder detection, but building
annotated datasets for this task can be challenging; moreover,
annotated data is very scarce for disorders other than depression.
Understanding the commonalities between certain disorders is also
important for clinicians who face the problem of shifting standards
of diagnosis. We propose that transfer learning with linguistic
features can be useful for approaching both the technical problem of
improving mental disorder detection in the context of data scarcity,
and the clinical problem of understanding the overlapping symptoms
between certain disorders. In this paper, we target four disorders:
depression, PTSD, anorexia and self-harm. We explore multi-aspect
transfer learning for detecting mental disorders from social media
texts, using deep learning models with multi-aspect representations
of language (including multiple types of interpretable linguistic
features). We explore different transfer learning strategies for
cross-disorder and cross-platform transfer, and show that transfer
learning can be effective for improving prediction performance for
disorders where little annotated data is available. We offer
insights into which linguistic features are the most useful vehicles
for transferring knowledge, through ablation experiments, as well as
error analysis.},
  url   = {https://aclanthology.org/2022.lrec-1.343}
}

@InProceedings{mubarak-EtAl:2022:LREC,
  author   = {Mubarak, Hamdy and Hassan, Sabit and Chowdhury,
Shammur Absar and Alam, Firoj},
  title   = {ArCovidVac: Analyzing Arabic Tweets About COVID-19
Vaccination},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3220--3230},
  abstract = {The emergence of the COVID-19 pandemic and the first

global infodemic have changed our lives in many different ways. We relied on social media to get the latest information about COVID-19 pandemic and at the same time to disseminate information. The content in social media consisted not only health related advice, plans, and informative news from policymakers, but also contains conspiracies and rumors. It became important to identify such information as soon as they are posted to make an actionable decision (e.g., debunking rumors, or taking certain measures for traveling). To address this challenge, we develop and publicly release the first largest manually annotated Arabic tweet dataset, ArCovidVac, for COVID-19 vaccination campaign, covering many countries in the Arab region. The dataset is enriched with different layers of annotation, including, (i) Informativeness more vs. less importance of the tweets); (ii) fine-grained tweet content types (e.g., advice, rumors, restriction, authenticate news/information); and (iii) stance towards vaccination (pro-vaccination, neutral, anti-vaccination). Further, we performed in-depth analysis of the data, exploring the popularity of different vaccines, trending hashtags, topics, and presence of offensiveness in the tweets. We studied the data for individual types of tweets and temporal changes in stance towards vaccine. We benchmarked the ArCovidVac dataset using transformer architectures for informativeness, content types, and stance detection.},
  url   = {https://aclanthology.org/2022.lrec-1.344}
}

@InProceedings{sakketou-EtAl:2022:LREC1,
  author   = {Sakketou, Flora and Plepi, Joan and Cervero, Riccardo and Geiss, Henri Jacques and Rosso, Paolo and Flek, Lucie},
  title   = {FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3231--3241},
  abstract = {Proactively identifying misinformation spreaders is an important step towards mitigating the impact of fake news on our society. In this paper, we introduce a new contemporary Reddit dataset for fake news spreader analysis, called FACTOID, monitoring political discussions on Reddit since the beginning of 2020. The dataset contains over 4K users with 3.4M Reddit posts, and includes, beyond the users' binary labels, also their fine-grained credibility level (very low to very high) and their political bias strength (extreme right to extreme left). As far as we are aware, this is the first fake news spreader dataset that simultaneously captures both the long-term context of users' historical posts and the interactions between them. To create the first benchmark on our data, we provide methods for identifying misinformation spreaders by utilizing the social connections between the users along with their psycho-linguistic features. We show that the users’ social

interactions can, on their own, indicate misinformation spreading, while the psycho-linguistic features are mostly informative in non-neural classification settings. In a qualitative analysis we observe that detecting affective mental processes correlates negatively with right-biased users, and that the openness to experience factor is lower for those who spread fake news.},
  url   = {https://aclanthology.org/2022.lrec-1.345}
}

@InProceedings{pritzen-EtAl:2022:LREC,
  author   = {Pritzen, Julia and Gref, Michael and Zühlke, Dietlind and Schmidt, Christoph Andreas},
  title   = {Multitask Learning for Grapheme-to-Phoneme Conversion of Anglicisms in German Speech Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3242--3249},
  abstract = {Anglicisms are a challenge in German speech recognition. Due to their irregular pronunciation compared to native German words, automatically generated pronunciation dictionaries often contain incorrect phoneme sequences for Anglicisms. In this work, we propose a multitask sequence-to-sequence approach for grapheme-to-phoneme conversion to improve the phonetization of Anglicisms. We extended a grapheme-to-phoneme model with a classification task to distinguish Anglicisms from native German words. With this approach, the model learns to generate different pronunciations depending on the classification result. We used our model to create supplementary Anglicism pronunciation dictionaries to be added to an existing German speech recognition model. Tested on a special Anglicism evaluation set, we improved the recognition of Anglicisms compared to a baseline model, reducing the word error rate by a relative 1 \% and the Anglicism error rate by a relative 3 \%. With our experiment, we show that multitask learning can help solving the challenge of Anglicisms in German speech recognition.},
  url   = {https://aclanthology.org/2022.lrec-1.346}
}

@InProceedings{plss-EtAl:2022:LREC,
  author   = {Plüss, Michel and Hürlimann, Manuela and Cuny, Marc and Stöckli, Alla and Kapotis, Nikolaos and Hartmann, Julia and Ulasik, Malgorzata Anna and Scheller, Christian and Schraner, Yanick and Jain, Amit and Deriu, Jan and Cieliebak, Mark and Vogel, Manfred},
  title   = {SDS-200: A Swiss German Speech to Standard German Text Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

publisher = {European Language Resources Association},
pages = {3250--3256},
abstract = {We present SDS-200, a corpus of Swiss German dialectal speech with Standard German text translations, annotated with dialect, age, and gender information of the speakers. The dataset allows for training speech translation, dialect recognition, and speech synthesis systems, among others. The data was collected using a web recording tool that is open to the public. Each participant was given a text in Standard German and asked to translate it to their Swiss German dialect before recording it. To increase the corpus quality, recordings were validated by other participants. The data consists of 200 hours of speech by around 4000 different speakers and covers a large part of the Swiss German dialect landscape. We release SDS-200 alongside a baseline speech translation model, which achieves a word error rate (WER) of 30.3 and a BLEU score of 53.1 on the SDS-200 test set. Furthermore, we use SDS-200 to fine-tune a pre-trained XLS-R model, achieving 21.6 WER and 64.0 BLEU.},
url = {https://aclanthology.org/2022.lrec-1.347}
}

@InProceedings{wu-EtAl:2022:LREC2,
author = {WU, Yaru and Hutin, Mathilde and Vasilescu, Ioana and Lamel, Lori and Adda-Decker, Martine},
title = {Extracting Linguistic Knowledge from Speech: A Study of Stop Realization in 5 Romance Languages},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3257--3263},
abstract = {This paper builds upon recent work in leveraging the corpora and tools originally used to develop speech technologies for corpus-based linguistic studies. We address the non-canonical realization of consonants in connected speech and we focus on voicing alternation phenomena of stops in 5 standard varieties of Romance languages (French, Italian, Spanish, Portuguese, Romanian). For these languages, both large scale corpora and speech recognition systems were available for the study. We use forced alignment with pronunciation variants and machine learning techniques to examine to what extent such frequent phenomena characterize languages and what are the most triggering factors. The results confirm that voicing alternations occur in all Romance languages. Automatic classification underlines that surrounding contexts and segment duration are recurring contributing factors for modeling voicing alternation. The results of this study also demonstrate the new role that machine learning techniques such as classification algorithms can play in helping to extract linguistic knowledge from speech and to suggest interesting research directions.},
url = {https://aclanthology.org/2022.lrec-1.348}
}

@InProceedings{lebourdais-EtAl:2022:LREC,
  author   = {Lebourdais, Martin and Tahon, Marie and LAURENT, Antoine and Meignier, Sylvain and Larcher, Anthony},
  title   = {Overlaps and Gender Analysis in the Context of Broadcast Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3264--3270},
  abstract = {Our main goal is to study the interactions between speakers according to their gender and role in broadcast media. In this paper, we propose an extensive study of gender and overlap annotations in various speech corpora mainly dedicated to diarisation or transcription tasks. We point out the issue of the heterogeneity of the annotation guidelines for both overlapping speech and gender categories. On top of that, we analyse how the speech content (casual speech, meetings, debate, interviews, etc.) impacts the distribution of overlapping speech segments. On a small dataset of 93 recordings from LCP French channel, we intend to characterise the interactions between speakers according to their gender. Finally, we propose a method which aims to highlight active speech areas in terms of interactions between speakers. Such a visualisation tool could improve the efficiency of qualitative studies conducted by researchers in human sciences.},
  url   = {https://aclanthology.org/2022.lrec-1.349}
}

@InProceedings{minh-EtAl:2022:LREC,
  author   = {Minh, Nguyen and Tran, Vu Hoang and Hoang, Vu and Ta, Huy Duc and Bui, Trung Huu and Truong, Steven Quoc Hung},
  title   = {ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {328--337},
  abstract = {Pre-trained language models have become crucial to achieving competitive results across many Natural Language Processing (NLP) problems. For monolingual pre-trained models in low-resource languages, the quantity has been significantly increased. However, most of them relate to the general domain, and there are limited strong baseline language models for domain-specific. We introduce ViHealthBERT, the first domain-specific pre-trained language model for Vietnamese healthcare. The performance of our model shows strong results while outperforming the general domain language models in all health-related datasets. Moreover, we also present Vietnamese datasets for the healthcare domain for two

tasks are Acronym Disambiguation (AD) and Frequently Asked Questions (FAQ) Summarization. We release our ViHealthBERT to facilitate future research and downstream application for Vietnamese NLP in domain-specific. Our dataset and code are available in https://github.com/demdecuong/vihealthbert.},
  url = {https://aclanthology.org/2022.lrec-1.35}
}

@InProceedings{uro-EtAl:2022:LREC,
  author = {Uro, Rémi and Doukhan, David and Rilliard, Albert and Larcher, Laetitia and Adgharouamane, Anissa-Claire and Tahon, Marie and Laurent, Antoine},
  title = {A Semi-Automatic Approach to Create Large Gender- and Age-Balanced Speaker Corpora: Usefulness of Speaker Diarization \& Identification.},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {3271--3280},
  abstract = {This paper presents a semi-automatic approach to create a diachronic corpus of voices balanced for speaker's age, gender, and recording period, according to 32 categories (2 genders, 4 age ranges and 4 recording periods). Corpora were selected at French National Institute of Audiovisual (INA) to obtain at least 30 speakers per category (a total of 960 speakers; only 874 have be found yet). For each speaker, speech excerpts were extracted from audiovisual documents using an automatic pipeline consisting of speech detection, background music and overlapped speech removal and speaker diarization, used to present clean speaker segments to human annotators identifying target speakers. This pipeline proved highly effective, cutting down manual processing by a factor of ten. Evaluation of the quality of the automatic processing and of the final output is provided. It shows the automatic processing compare to up-to-date process, and that the output provides high quality speech for most of the selected excerpts. This method is thus recommendable for creating large corpora of known target speakers.},
  url = {https://aclanthology.org/2022.lrec-1.350}
}

@InProceedings{scholman-EtAl:2022:LREC2,
  author = {Scholman, Merel and Dong, Tianai and Yung, Frances and Demberg, Vera},
  title = {DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {3281--3290},

  abstract = {We present DiscoGeM, a crowdsourced corpus of 6,505 implicit discourse relations from three genres: political speech, literature, and encyclopedic texts. Each instance was annotated by 10 crowd workers. Various label aggregation methods were explored to evaluate how to obtain a label that best captures the meaning inferred by the crowd annotators. The results show that a significant proportion of discourse relations in DiscoGeM are ambiguous and can express multiple relation senses. Probability distribution labels better capture these interpretations than single labels. Further, the results emphasize that text genre crucially affects the distribution of discourse relations, suggesting that genre should be included as a factor in automatic relation classification. We make available the newly created DiscoGeM corpus, as well as the dataset with all annotator-level labels. Both the corpus and the dataset can facilitate a multitude of applications and research purposes, for example to function as training data to improve the performance of automatic discourse relation parsers, as well as facilitate research into non-connective signals of discourse relations.},
  url   = {https://aclanthology.org/2022.lrec-1.351}
}

@InProceedings{hautlijanisz-EtAl:2022:LREC,
  author   = {Hautli-Janisz, Annette and Kikteva, Zlata and Siskou, Wassiliki and Gorska, Kamila and Becker, Ray and Reed, Chris},
  title   = {QT30: A Corpus of Argument and Conflict in Broadcast Debate},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3291--3300},
  abstract = {Broadcast political debate is a core pillar of democracy: it is the public’s easiest access to opinions that shape policies and enables the general public to make informed choices. With QT30, we present the largest corpus of analysed dialogical argumentation ever created (19,842 utterances, 280,000 words) and also the largest corpus of analysed broadcast political debate to date, using 30 episodes of BBC’s ‘Question Time’ from 2020 and 2021. Question Time is the prime institution in UK broadcast political debate and features questions from the public on current political issues, which are responded to by a weekly panel of five figures of UK politics and society. QT30 is highly argumentative and combines language of well-versed political rhetoric with direct, often combative, justification-seeking of the general public. QT30 is annotated with Inference Anchoring Theory, a framework well-known in argument mining, which encodes the way arguments and conflicts are created and reacted to in dialogical settings. The resource is freely available at http://corpora.aifdb.org/qt30.},
  url   = {https://aclanthology.org/2022.lrec-1.352}
}

@InProceedings{falk-lapesa:2022:LREC,
  author   = {Falk, Neele and Lapesa, Gabriella},
  title   = {Scaling up Discourse Quality Annotation for Political Science},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3301--3318},
  abstract = {The empirical quantification of the quality of a contribution to a political discussion is at the heart of deliberative theory, the subdiscipline of political science which investigates decision-making in deliberative democracy. Existing annotation on deliberative quality is time-consuming and carried out by experts, typically resulting in small datasets which also suffer from strong class imbalance. Scaling up such annotations with automatic tools is desirable, but very challenging. We take up this challenge and explore different strategies to improve the prediction of deliberative quality dimensions (justification, common good, interactivity, respect) in a standard dataset. Our results show that simple data augmentation techniques successfully alleviate data imbalance. Classifiers based on linguistic features (textual complexity and sentiment/polarity) and classifiers integrating argument quality annotations (from the argument mining community in NLP) were consistently outperformed by transformer-based models, with or without data augmentation.},
  url   = {https://aclanthology.org/2022.lrec-1.353}
}

@InProceedings{anthonio-sauer-roth:2022:LREC,
  author   = {Anthonio, Talita and Sauer, Anna and Roth, Michael},
  title   = {Clarifying Implicit and Underspecified Phrases in Instructional Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3319--3330},
  abstract = {Natural language inherently consists of implicit and underspecified phrases, which represent potential sources of misunderstanding. In this paper, we present a data set of such phrases in English from instructional texts together with multiple possible clarifications. Our data set, henceforth called CLAIRE, is based on a corpus of revision histories from wikiHow, from which we extract human clarifications that resolve an implicit or underspecified phrase. We show how language modeling can be used to generate alternate clarifications, which may or may not be compatible with the human clarification. Based on plausibility

judgements for each clarification, we define the task of distinguishing between plausible and implausible clarifications. We provide several baseline models for this task and analyze to what extent different clarifications represent multiple readings as a first step to investigate misunderstandings caused by implicit/ underspecified language in instructional texts.},
  url   = {https://aclanthology.org/2022.lrec-1.354}
}

@InProceedings{buzanov-EtAl:2022:LREC,
  author   = {Buzanov, Anton and Bychkova, Polina and Molchanova, Arina and Postnikova, Anna and Ryzhova, Daria},
  title   = {Multilingual Pragmaticon: Database of Discourse Formulae},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3331--3336},
  abstract = {The paper presents a multilingual database aimed to be used as a tool for typological analysis of response constructions called discourse formulae (DF), cf. English `No way!` or French `Ça va!` (~ `all right'). The two primary qualities that make DF of theoretical interest for linguists are their idiomaticity and the special nature of their meanings (cf. consent, refusal, negation), determined by their dialogical function. The formal and semantic structures of these items are language-specific. Compiling a database with DF from various languages would help estimate the diversity of DF in both of these aspects, and, at the same time, establish some frequently occurring patterns. The DF in the database are accompanied with glosses and assigned with multiple tags, such as pragmatic function, additional semantics, the illocutionary type of the context, etc. As a starting point, Russian, Serbian and Slovene DF are included into the database. This data already shows substantial grammatical and lexical variability.},
  url   = {https://aclanthology.org/2022.lrec-1.355}
}

@InProceedings{stankovi-EtAl:2022:LREC,
  author   = {Stanković, Ranka and Krstev, Cvetana and Šandrih Todorović, Branislava and Vitas, Dusko and Skoric, Mihailo and Ikonić Nešić, Milica},
  title   = {Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3337--3345},
  abstract = {In this paper we present the Serbian part of the

ELTeC multilingual corpus of novels written in the time period 1840-1920. The corpus is being built in order to test various distant reading methods and tools with the aim of re-thinking the European literary history. We present the various steps that led to the production of the Serbian sub-collection: the novel selection and retrieval, text preparation, structural annotation, POS-tagging, lemmatization and named entity recognition. The Serbian sub-collection was published on different platforms in order to make it freely available to various users. Several use examples show that this sub-collection is usefull for both close and distant reading approaches.},
  url   = {https://aclanthology.org/2022.lrec-1.356}
}

@InProceedings{reiter-EtAl:2022:LREC,
  author   = {Reiter, Nils and Sieker, Judith and Guhr, Svenja and Gius, Evelyn and Zarrieß, Sina},
  title   = {Exploring Text Recombination for Automatic Narrative Level Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3346--3353},
  abstract = {Automatizing the process of understanding the global narrative structure of long texts and stories is still a major challenge for state-of-the-art natural language understanding systems, particularly because annotated data is scarce and existing annotation workflows do not scale well to the annotation of complex narrative phenomena. In this work, we focus on the identification of narrative levels in texts corresponding to stories that are embedded in stories. Lacking sufficient pre-annotated training data, we explore a solution to deal with data scarcity that is common in machine learning: the automatic augmentation of an existing small data set of annotated samples with the help of data synthesis. We present a workflow for narrative level detection, that includes the operationalization of the task, a model, and a data augmentation protocol for automatically generating narrative texts annotated with breaks between narrative levels. Our experiments suggest that narrative levels in long text constitute a challenging phenomenon for state-of-the-art NLP models, but generating training data synthetically does improve the prediction results considerably.},
  url   = {https://aclanthology.org/2022.lrec-1.357}
}

@InProceedings{bawden-EtAl:2022:LREC,
  author   = {Bawden, Rachel and Poinhos, Jonathan and Kogkitsidou, Eleni and Gambette, Philippe and Sagot, Benoît and Gabay, Simon},
  title   = {Automatic Normalisation of Early Modern French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3354--3366},
  abstract = {Spelling normalisation is a useful step in the study and analysis of historical language texts, whether it is manual analysis by experts or automatic analysis using downstream natural language processing (NLP) tools. Not only does it help to homogenise the variable spelling that often exists in historical texts, but it also facilitates the use of off-the-shelf contemporary NLP tools, if contemporary spelling conventions are used for normalisation. We present FREEMnorm, a new benchmark for the normalisation of Early Modern French (from the 17th century) into contemporary French and provide a thorough comparison of three different normalisation methods: ABA, an alignment-based approach and MT-approaches, (both statistical and neural), including extensive parameter searching, which is often missing in the normalisation literature.},
  url   = {https://aclanthology.org/2022.lrec-1.358}
}

@InProceedings{gabay-EtAl:2022:LREC,
  author   = {Gabay, Simon and Ortiz Suarez, Pedro and BARTZ, Alexandre and Chagué, Alix and Bawden, Rachel and Gambette, Philippe and Sagot, Benoît},
  title   = {From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3367--3374},
  abstract = {anguage models for historical states of language are becoming increasingly important to allow the optimal digitisation and analysis of old textual sources. Because these historical states are at the same time more complex to process and more scarce in the corpora available, this paper presents recent efforts to overcome this difficult situation. These efforts include producing a corpus, creating the model, and evaluating it with an NLP task currently used by scholars in other ongoing projects.},
  url   = {https://aclanthology.org/2022.lrec-1.359}
}

@InProceedings{igamberdiev-habernal:2022:LREC,
  author   = {Igamberdiev, Timour and Habernal, Ivan},
  title   = {Privacy-Preserving Graph Convolutional Networks for Text Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {338--350},
  abstract = {Graph convolutional networks (GCNs) are a powerful architecture for representation learning on documents that naturally occur as graphs, e.g., citation or social networks. However, sensitive personal information, such as documents with people's profiles or relationships as edges, are prone to privacy leaks, as the trained model might reveal the original input. Although differential privacy (DP) offers a well-founded privacy-preserving framework, GCNs pose theoretical and practical challenges due to their training specifics. We address these challenges by adapting differentially-private gradient-based training to GCNs and conduct experiments using two optimizers on five NLP datasets in two languages. We propose a simple yet efficient method based on random graph splits that not only improves the baseline privacy bounds by a factor of 2.7 while retaining competitive F1 scores, but also provides strong privacy guarantees of epsilon = 1.0. We show that, under certain modeling choices, privacy-preserving GCNs perform up to 90\% of their non-private variants, while formally guaranteeing strong privacy measures.},
  url   = {https://aclanthology.org/2022.lrec-1.36}
}

@InProceedings{heyns-vanzaanen:2022:LREC,
  author   = {Heyns, Nuette and van Zaanen, Menno},
  title   = {Detecting Multiple Transitions in Literary Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3375--3381},
  abstract = {Identifying the high level structure of texts provides important information when performing distant reading analysis. The structure of texts is not necessarily linear, as transitions, such as changes in the scenery or flashbacks, can be present. As a first step in identifying this structure, we aim to identify transitions in texts. Previous work (Heyns and van Zaanen, 2021) proposed a system that can successfully identify one transition in literary texts. The text is split in snippets and LDA is applied, resulting in a sequence of topics. A transition is introduced at the point that separates the topics (before and after the point) best. In this article, we extend the existing system such that it can detect multiple transitions. Additionally, we introduce a new system that inherently handles multiple transitions in texts. The new system also relies on LDA information, but is more robust than the previous system. We apply these systems to texts with known transitions (as they are constructed by concatenating text snippets stemming from different source texts) and evaluation both systems on texts with one transition and texts with two transitions. As both systems rely on LDA to identify transitions between snippets, we also show the impact of varying the number of LDA topics on the results as well. The new system consistently outperforms the

previous system, not only on texts with multiple transitions, but also on single boundary texts.},
  url   = {https://aclanthology.org/2022.lrec-1.360}
}

@InProceedings{escribano-EtAl:2022:LREC,
  author   = {Escribano, Nayla and Gonzalez, Jon Ander and Orbegozo-Terradillos, Julen and Larrondo-Ureta, Ainara and Peña-Fernández, Simón and Perez-de-Viñaspre, Olatz and Agerri, Rodrigo},
  title   = {BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3382--3390},
  abstract = {Parliamentary transcripts provide a valuable resource to understand the reality and know about the most important facts that occur over time in our societies. Furthermore, the political debates captured in these transcripts facilitate research on political discourse from a computational social science perspective. In this paper we release the first version of a newly compiled corpus from Basque parliamentary transcripts. The corpus is characterized by heavy Basque-Spanish code-switching, and represents an interesting resource to study political discourse in contrasting languages such as Basque and Spanish. We enrich the corpus with metadata related to relevant attributes of the speakers and speeches (language, gender, party...) and process the text to obtain named entities and lemmas. The obtained metadata is then used to perform a detailed corpus analysis which provides interesting insights about the language use of the Basque political representatives across time, parties and gender.},
  url   = {https://aclanthology.org/2022.lrec-1.361}
}

@InProceedings{poppek-masloch-kiss:2022:LREC,
  author   = {Poppek, Johanna M. and Masloch, Simon and Kiss, Tibor},
  title   = {GerEO: A Large-Scale Resource on the Syntactic Distribution of German Experiencer-Object Verbs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3391--3397},
  abstract = {Although studied for several decades, the syntactic properties of experiencer-object (EO) verbs are still under discussion, while most analyses are not supported by substantial corpus data. With GerEO, we intend to fill this lacuna for German

EO-verbs by presenting a large-scale database of more than 10,000 examples for 64 verbs (up to 200 per verb) from a newspaper corpus annotated for several syntactic and semantic features relevant for their analysis, including the overall syntactic construction, the semantic stimulus type, and the form of a possible stimulus preposition, i.e. a preposition heading a PP that indicates (a part/aspect of) the stimulus. Non-psych occurrences of the verbs are not excluded from the database but marked as such to make a comparison possible. Data of this kind can be used to develop and test theoretical hypotheses on the properties of EO-verbs, aid in the construction of experiments as well as provide training and test data for AI systems.},
  url   = {https://aclanthology.org/2022.lrec-1.362}
}

@InProceedings{nambanoorkunnath-EtAl:2022:LREC,
  author   = {Nambanoor Kunnath, Suchetha and Stauber, Valentin and Wu, Ronin and Pride, David and Botev, Viktor and Knoth, Petr},
  title   = {ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3398--3406},
  abstract = {Classifying citations according to their purpose and importance is a challenging task that has gained considerable interest in recent years. This interest has been primarily driven by the need to create more transparent, efficient, merit-based reward systems in academia; a system that goes beyond simple bibliometric measures and considers the semantics of citations. Such systems that quantify and classify the influence of citations can act as edges that link knowledge nodes to a graph and enable efficient knowledge discovery. While a number of researchers have experimented with a variety of models, these experiments are typically limited to single-domain applications and the resulting models are hardly comparable. Recently, two Citation Context Classification (3C) shared tasks (at WOSP2020 and SDP2021) created the first benchmark enabling direct comparison of citation classification approaches, revealing the crucial impact of supplementary data on the performance of models. Reflecting from the findings of these shared tasks, we are releasing a new multi-disciplinary dataset, ACT2, an extended SDP 3C shared task dataset. This modified corpus has annotations for both citation function and importance classes newly enriched with supplementary contextual and non-contextual feature sets the selection of which follows from the lists of features used by the more successful teams in these shared tasks. Additionally, we include contextual features for cited papers (e.g. Abstract of the cited paper), which most existing datasets lack, but which have a lot of potential to improve results. We describe the methodology used for feature extraction and the challenges involved in the

process. The feature enriched ACT2 dataset is available at https://
github.com/oacore/ACT2.},
  url   = {https://aclanthology.org/2022.lrec-1.363}
}

@InProceedings{bunt-EtAl:2022:LREC,
  author   = {Bunt, Harry and Amblard, Maxime and Bos, Johan
and Fort, Karën and Guillaume, Bruno and de Groote, Philippe
and Li, Chuyuan and Ludmann, Pierre and Musiol, Michel and
Pavlova, Siyana and Perrier, Guy and Pogodalla, Sylvain},
  title   = {Quantification Annotation in ISO 24617-12, Second
Draft},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3407--3416},
  abstract = {This paper describes the continuation of a project
that aims at establishing an interoperable annotation schema for
quantification phenomena as part of the ISO suite of standards for
semantic annotation, known as the Semantic Annotation Framework.
After a break, caused by the Covid-19 pandemic, the project was
relaunched in early 2022 with a second working draft of an
annotation scheme, which is discussed in this paper. Keywords:
semantic annotation, quantification, interoperability, annotation
schema, ISO standard},
  url   = {https://aclanthology.org/2022.lrec-1.364}
}

@InProceedings{mujadia-sharma:2022:LREC,
  author   = {Mujadia, Vandan and Sharma, Dipti},
  title   = {The LTRC Hindi-Telugu Parallel Corpus},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3417--3424},
  abstract = {We present the Hindi-Telugu Parallel Corpus of
different technical domains such as Natural Science, Computer
Science, Law and Healthcare along with the General domain. The
qualitative corpus consists of 700K parallel sentences of which 535K
sentences were created using multiple methods such as extract, align
and review of Hindi-Telugu corpora, end-to-end human translation,
iterative back-translation driven post-editing and around 165K
parallel sentences were collected from available sources in the
public domain. We present the comparative assessment of created
parallel corpora for representativeness and diversity. The corpus
has been pre-processed for machine translation, and we trained a
neural machine translation system using it and report state-of-the-
art baseline results on the developed development set over multiple

domains and on available benchmarks. With this, we define a new task on Domain Machine Translation for low resource language pairs such as Hindi and Telugu. The developed corpus (535K) is freely available for non-commercial research and to the best of our knowledge, this is the well curated, largest, publicly available domain parallel corpus for Hindi-Telugu.},
  url   = {https://aclanthology.org/2022.lrec-1.365}
}

@InProceedings{rani-mccrae-fransen:2022:LREC,
  author   = {Rani, Priya and McCrae, John P. and Fransen, Theodorus},
  title   = {MHE: Code-Mixed Corpora for Similar Language Identification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3425--3433},
  abstract = {This paper introduces a new Magahi-Hindi-English (MHE) code-mixed data-set for similar language identification (SMLID), where Magahi is a less-resourced minority language. This corpus provides a language id at two levels: word and sentence. This data-set is the first Magahi-Hindi-English code-mixed data-set for similar language identification task. Furthermore, we will discuss the complexity of the data-set and provide a few baselines for the language identification task.},
  url   = {https://aclanthology.org/2022.lrec-1.366}
}

@InProceedings{lerner-EtAl:2022:LREC,
  author   = {Lerner, Paul and Bergoënd, Juliette and Guinaudeau, Camille and Bredin, Hervé and Maurice, Benjamin and Lefevre, Sharleyne and Bouteiller, Martin and Berhe, Aman and Galmant, Léo and Yin, Ruiqing and Barras, Claude},
  title   = {Bazinga! A Dataset for Multi-Party Dialogues Structuring},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3434--3441},
  abstract = {We introduce a dataset built around a large collection of TV (and movie) series. Those are filled with challenging multi-party dialogues. Moreover, TV series come with a very active fan base that allows the collection of metadata and accelerates annotation. With 16 TV and movie series, Bazinga! amounts to 400+ hours of speech and 8M+ tokens, including 500K+ tokens annotated with the speaker, addressee, and entity linking information. Along with the dataset, we also provide a baseline for

speaker diarization, punctuation restoration, and person entity recognition. The results demonstrate the difficulty of the tasks and of transfer learning from models trained on mono-speaker audio or written text, which is more widely available. This work is a step towards better multi-party dialogue structuring and understanding. Bazinga! is available at hf.co/bazinga. Because (a large) part of Bazinga! is only partially annotated, we also expect this dataset to foster research towards self- or weakly-supervised learning methods.},
  url   = {https://aclanthology.org/2022.lrec-1.367}
}

@InProceedings{ntogramatzis-EtAl:2022:LREC,
  author   = {Ntogramatzis, Alexandros Fotios and Gradou, Anna and Petasis, Georgios and Kokol, Marko},
  title   = {The Ellogon Web Annotation Tool: Annotating Moral Values and Arguments},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3442--3450},
  abstract = {In this paper, we present the Ellogon Web Annotation Tool. It is a collaborative, web-based annotation tool built upon the Ellogon infrastructure offering an improved user experience and adaptability to various annotation scenarios by making good use of the latest design practices and web development frameworks. Being in development for many years, this paper describes its current architecture, along with the recent modifications that extend the existing functionalities and the new features that were added. The new version of the tool offers document analytics, annotation inspection and comparison features, a modern UI, and formatted text import (e.g. TEI XML documents, rendered with simple markup). We present two use cases that serve as two examples of different annotation scenarios to demonstrate the new functionalities. An appropriate (user-supplied, XML-based) annotation schema is used for each scenario. The first schema contains the relevant components for representing concepts, moral values, and ideas. The second includes all the necessary elements for annotating argumentative units in a document and their binary relations.},
  url   = {https://aclanthology.org/2022.lrec-1.368}
}

@InProceedings{jones-EtAl:2022:LREC,
  author   = {Jones, Karen and Walker, Kevin and Caruso, Christopher and Wright, Jonathan and Strassel, Stephanie},
  title   = {WeCanTalk: A New Multi-language, Multi-modal Resource for Speaker Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3451--3456},
abstract = {The WeCanTalk (WCT) Corpus is a new multi-language, multi-modal resource for speaker recognition. The corpus contains Cantonese, Mandarin and English telephony and video speech data from over 200 multilingual speakers located in Hong Kong. Each speaker contributed at least 10 telephone conversations of 8-10 minutes' duration collected via a custom telephone platform based in Hong Kong. Speakers also uploaded at least 3 videos in which they were both speaking and visible, along with one selfie image. At least half of the calls and videos for each speaker were in Cantonese, while their remaining recordings featured one or more different languages. Both calls and videos were made in a variety of noise conditions. All speech and video recordings were audited by experienced multilingual annotators for quality including presence of the expected language and for speaker identity. The WeCanTalk Corpus has been used to support the NIST 2021 Speaker Recognition Evaluation and will be published in the LDC catalog.},
url = {https://aclanthology.org/2022.lrec-1.369}
}

@InProceedings{alghamdi-liang-zhang:2022:LREC,
author = {Alghamdi, Reem and Liang, Zhenwen and Zhang, Xiangliang},
title = {ArMATH: a Dataset for Solving Arabic Math Word Problems},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {351--362},
abstract = {This paper studies solving Arabic Math Word Problems by deep learning. A Math Word Problem (MWP) is a text description of a mathematical problem that can be solved by deriving a math equation to reach the answer. Effective models have been developed for solving MWPs in English and Chinese. However, Arabic MWPs are rarely studied. This paper contributes the first large-scale dataset for Arabic MWPs, which contains 6,000 samples of primary-school math problems, written in Modern Standard Arabic (MSA). Arabic MWP solvers are then built with deep learning models and evaluated on this dataset. In addition, a transfer learning model is built to let the high-resource Chinese MWP solver promote the performance of the low-resource Arabic MWP solver. This work is the first to use deep learning methods to solve Arabic MWP and the first to use transfer learning to solve MWP across different languages. The transfer learning enhanced solver has an accuracy of 74.15\%, which is 3\% higher than the solver without using transfer learning. We make the dataset and solvers available in public for encouraging more research of Arabic MWPs: https://github.com/reem-codes/ArMATH},
url = {https://aclanthology.org/2022.lrec-1.37}
}

@InProceedings{bajeti-declerck:2022:LREC,
  author   = {Bajčetić, Lenka and Declerck, Thierry},
  title   = {Using Wiktionary to Create Specialized Lexical Resources and Datasets},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3457--3460},
  abstract = {This paper describes an approach aiming at utilizing Wiktionary data for creating specialized lexical datasets which can be used for enriching other lexical (semantic) resources or for generating datasets that can be used for evaluating or improving NLP tasks, like Word Sense Disambiguation, Word-in-Context challenges, or Sense Linking across lexicons and dictionaries. We have focused on Wiktionary data about pronunciation information in English, and grammatical number and grammatical gender in German.},
  url   = {https://aclanthology.org/2022.lrec-1.370}
}

@InProceedings{zhang-wilson-mitra:2022:LREC,
  author   = {Zhang, Nan and Wilson, Shomir and Mitra, Prasenjit},
  title   = {STAPI: An Automatic Scraper for Extracting Iterative Title-Text Structure from Web Documents},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3461--3470},
  abstract = {Formal documents often are organized into sections of text, each with a title, and extracting this structure remains an under-explored aspect of natural language processing. This iterative title-text structure is valuable data for building models for headline generation and section title generation, but there is no corpus that contains web documents annotated with titles and prose texts. Therefore, we propose the first title-text dataset on web documents that incorporates a wide variety of domains to facilitate downstream training. We also introduce STAPI (Section Title And Prose text Identifier), a two-step system for labeling section titles and prose text in HTML documents. To filter out unrelated content like document footers, its first step involves a filter that reads HTML documents and proposes a set of textual candidates. In the second step, a typographic classifier takes the candidates from the filter and categorizes each one into one of the three pre-defined classes (title, prose text, and miscellany). We show that STAPI significantly outperforms two baseline models in terms of title-text identification. We release our dataset along with a web application to facilitate supervised and semi-supervised training in

this domain.},
  url   = {https://aclanthology.org/2022.lrec-1.371}
}

@InProceedings{horvth-EtAl:2022:LREC,
  author   = {Horváth, Péter and Kundráth, Péter and Indig, Balázs and Fellegi, Zsófia and Szlávich, Eszter and Bajzát, Tímea Borbála and Sárközi-Lindner, Zsófia and Vida, Bence and Karabulut, Aslihan and Timári, Mária and Palkó, Gábor},
  title   = {ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3471--3478},
  abstract = {ELTE Poetry Corpus is a database that stores canonical Hungarian poetry with automatically generated annotations of the poems' structural units, grammatical features and sound devices, i.e. rhyme patterns, rhyme pairs, rhythm, alliterations and the main phonological features of words. The corpus has an open access online query tool with several search functions. The paper presents the main stages of the annotation process and the tools used for each stage. The TEI XML format of the different versions of the corpus, each of which contains an increasing number of annotation layers, is presented as well. We have also specified our own XML format for the corpus, slightly different from TEI, in order to make it easier and faster to execute queries on the corpus. We discuss the results of a manual evaluation of the quality of automatic annotation of rhythm, as well as the results of an automatic evaluation of different rule sets used for the automatic annotation of rhyme patterns. Finally, the paper gives an overview of the main functions of the online query tool developed for the corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.372}
}

@InProceedings{sharma-mishra-sharma:2022:LREC,
  author   = {Sharma, Harshita and Mishra, Pruthwik and Sharma, Dipti},
  title   = {HAWP: a Dataset for Hindi Arithmetic Word Problem Solving},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3479--3490},
  abstract = {Word Problem Solving remains a challenging and interesting task in NLP. A lot of research has been carried out to solve different genres of word problems with various complexity

levels in recent years. However, most of the publicly available datasets and work has been carried out for English. Recently there has been a surge in this area of word problem solving in Chinese with the creation of large benchmark datastes. Apart from these two languages, labeled benchmark datasets for low resource languages are very scarce. This is the first attempt to address this issue for any Indian Language, especially Hindi. In this paper, we present HAWP (Hindi Arithmetic Word Problems), a dataset consisting of 2336 arithmetic word problems in Hindi. We also developed baseline systems for solving these word problems. We also propose a new evaluation technique for word problem solvers taking equation equivalence into account.},
  url   = {https://aclanthology.org/2022.lrec-1.373}
}

@InProceedings{osenova-EtAl:2022:LREC,
  author   = {Osenova, Petya and Simov, Kiril and Marinova, Iva and Berbatova, Melania},
  title   = {The Bulgarian Event Corpus: Overview and Initial NER Experiments},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3491--3499},
  abstract = {The paper describes the Bulgarian Event Corpus (BEC). The annotation scheme is based on CIDOC-CRM ontology and on the English Framenet, adjusted for our task. It includes two main layers: named entities and events with their roles. The corpus is multi-domain and mainly oriented towards Social Sciences and Humanities (SSH). It will be used for: extracting knowledge and making it available through the Bulgaria-centric Knowledge Graph; further developing an annotation scheme that handles multiple domains in SSH; training automatic modules for the most important knowledge-based tasks, such as domain-specific and nested NER, NEL, event detection and profiling. Initial experiments were conducted on standard NER task due to complexity of the dataset and the rich NE annotation scheme. The results are promising with respect to some labels and give insights on handling better other ones. These experiments serve also as error detection modules that would help us in scheme re-design. They are a basis for further and more complex tasks, such as nested NER, NEL and event detection.},
  url   = {https://aclanthology.org/2022.lrec-1.374}
}

@InProceedings{yao-EtAl:2022:LREC,
  author   = {Yao, Bingsheng and Joseph, Ethan and Lioanag, Julian and Si, Mei},
  title   = {A Corpus for Commonsense Inference in Story Cloze Test},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3500--3508},
  abstract = {The Story Cloze Test (SCT) is designed for training and evaluating machine learning algorithms for narrative understanding and inferences. The SOTA models can achieve over 90\% accuracy on predicting the last sentence. However, it has been shown that high accuracy can be achieved by merely using surface-level features. We suspect these models may not {\it truly} understand the story. Based on the SCT dataset, we constructed a human-labeled and human-verified commonsense knowledge inference dataset. Given the first four sentences of a story, we asked crowd-source workers to choose from four types of narrative inference for deciding the ending sentence and which sentence contributes most to the inference. We accumulated data on 1871 stories, and three human workers labeled each story. Analysis of the intra-category and inter-category agreements show a high level of consensus. We present two new tasks for predicting the narrative inference categories and contributing sentences. Our results show that transformer-based models can reach SOTA performance on the original SCT task using transfer learning but don't perform well on these new and more challenging tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.375}
}

@InProceedings{ekgren-EtAl:2022:LREC,
  author   = {Ekgren, Ariel and Cuba Gyllensten, Amaru and Gogoulou, Evangelia and Heiman, Alice and Verlinden, Severine and Öhman, Joey and Carlsson, Fredrik and Sahlgren, Magnus},
  title   = {Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3509--3518},
  abstract = {We present GTP-SW3, a 3.5 billion parameter autoregressive language model, trained on a newly created 100 GB Swedish corpus. This paper provides insights with regards to data collection and training, while highlights the challenges of proper model evaluation. The results of quantitive evaluation through perplexity indicate that GPT-SW3 is a competent model in comparison with existing autoregressive models of similar size. Additionally, we perform an extensive prompting study which reveals the good text generation capabilities of GTP-SW3.},
  url   = {https://aclanthology.org/2022.lrec-1.376}
}

@InProceedings{popescubelis-EtAl:2022:LREC,
  author   = {Popescu-Belis, Andrei and Atrio, Àlex and Minder,

Valentin and Xanthos, Aris and Luthier, Gabriel and Mattei, Simon and Rodriguez, Antonio},
title = {Constrained Language Models for Interactive Poem Generation},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3519--3529},
abstract = {This paper describes a system for interactive poem generation, which combines neural language models (LMs) for poem generation with explicit constraints that can be set by users on form, topic, emotion, and rhyming scheme. LMs cannot learn such constraints from the data, which is scarce with respect to their needs even for a well-resourced language such as French. We propose a method to generate verses and stanzas by combining LMs with rule-based algorithms, and compare several approaches for adjusting the words of a poem to a desired combination of topics or emotions. An approach to automatic rhyme setting using a phonetic dictionary is proposed as well. Our system has been demonstrated at public events, and log analysis shows that users found it engaging.},
url = {https://aclanthology.org/2022.lrec-1.377}
}

@InProceedings{lee-EtAl:2022:LREC2,
author = {Lee, Huije and NA, Young Ju and Song, Hoyun and Shin, Jisu and Park, Jong},
title = {ELF22: A Context-based Counter Trolling Dataset to Combat Internet Trolls},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3530--3541},
abstract = {Online trolls increase social costs and cause psychological damage to individuals. With the proliferation of automated accounts making use of bots for trolling, it is difficult for targeted individual users to handle the situation both quantitatively and qualitatively. To address this issue, we focus on automating the method to counter trolls, as counter responses to combat trolls encourage community users to maintain ongoing discussion without compromising freedom of expression. For this purpose, we propose a novel dataset for automatic counter response generation. In particular, we constructed a pair-wise dataset that includes troll comments and counter responses with labeled response strategies, which enables models fine-tuned on our dataset to generate responses by varying counter responses according to the specified strategy. We conducted three tasks to assess the effectiveness of our dataset and evaluated the results through both automatic and human evaluation. In human evaluation, we demonstrate

that the model fine-tuned with our dataset shows a significantly improved performance in strategy-controlled sentence generation.},
  url   = {https://aclanthology.org/2022.lrec-1.378}
}

@InProceedings{ampomah-EtAl:2022:LREC,
  author   = {Ampomah, Isaac and Burton, James and Enshaei, Amir and Al Moubayed, Noura},
  title   = {Generating Textual Explanations for Machine Learning Models Performance: A Table-to-Text Task},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3542--3551},
  abstract = {Numerical tables are widely employed to communicate or report the classification performance of machine learning (ML) models with respect to a set of evaluation metrics. For non-experts, domain knowledge is required to fully understand and interpret the information presented by numerical tables. This paper proposes a new natural language generation (NLG) task where neural models are trained to generate textual explanations, analytically describing the classification performance of ML models based on the metrics' scores reported in the tables. Presenting the generated texts along with the numerical tables will allow for a better understanding of the classification performance of ML models. We constructed a dataset comprising numerical tables paired with their corresponding textual explanations written by experts to facilitate this NLG task. Experiments on the dataset are conducted by fine-tuning pre-trained language models (T5 and BART) to generate analytical textual explanations conditioned on the information in the tables. Furthermore, we propose a neural module, Metrics Processing Unit (MPU), to improve the performance of the baselines in terms of correctly verbalising the information in the corresponding table. Evaluation and analysis conducted indicate, that exploring pre-trained models for data-to-text generation leads to better generalisation performance and can produce high-quality textual explanations.},
  url   = {https://aclanthology.org/2022.lrec-1.379}
}

@InProceedings{winter-EtAl:2022:LREC,
  author   = {Winter, Benjamin and Rosero, Alexei Figueroa and Löser, Alexander and Gers, Felix Alexander and Siu, Amy},
  title   = {KIMERA: Injecting Domain Knowledge into Vacant Transformer Heads},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {363--373},
  abstract = {Training transformer language models requires vast amounts of text and computational resources. This drastically limits the usage of these models in niche domains for which they are not optimized, or where domain-specific training data is scarce. We focus here on the clinical domain because of its limited access to training data in common tasks, while structured ontological data is often readily available. Recent observations in model compression of transformer models show optimization potential in improving the representation capacity of attention heads. We propose KIMERA (Knowledge Injection via Mask Enforced Retraining of Attention) for detecting, retraining and instilling attention heads with complementary structured domain knowledge. Our novel multi-task training scheme effectively identifies and targets individual attention heads that are least useful for a given downstream task and optimizes their representation with information from structured data. KIMERA generalizes well, thereby building the basis for an efficient fine-tuning. KIMERA achieves significant performance boosts on seven datasets in the medical domain in Information Retrieval and Clinical Outcome Prediction settings. We apply KIMERA to BERT-base to evaluate the extent of the domain transfer and also improve on the already strong results of BioBERT in the clinical domain.},
  url   = {https://aclanthology.org/2022.lrec-1.38}
}

@InProceedings{krjanec-edhi-demberg:2022:LREC,
  author   = {Škrjanec, Iza and Edhi, Muhammad Salman and Demberg, Vera},
  title   = {Barch: an English Dataset of Bar Chart Summaries},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3552--3560},
  abstract = {We present Barch, a new English dataset of human-written summaries describing bar charts. This dataset contains 47 charts based on a selection of 18 topics. Each chart is associated with one of the four intended messages expressed in the chart title. Using crowdsourcing, we collected around 20 summaries per chart, or one thousand in total. The text of the summaries is aligned with the chart data as well as with analytical inferences about the data drawn by humans. Our datasets is one of the first to explore the effect of intended messages on the data descriptions in chart summaries. Additionally, it lends itself well to the task of training data-driven systems for chart-to-text generation. We provide results on the performance of state-of-the-art neural generation models trained on this dataset and discuss the strengths and shortcomings of different models.},
  url   = {https://aclanthology.org/2022.lrec-1.380}
}

@InProceedings{martinc-EtAl:2022:LREC,
  author   = {Martinc, Matej and Montariol, Syrielle and
Pivovarova, Lidia and Zosa, Elaine},
  title   = {Effectiveness of Data Augmentation and Pretraining
for Improving Neural Headline Generation in Low-Resource Settings},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3561--3570},
  abstract = {We tackle the problem of neural headline generation
in a low-resource setting, where only limited amount of data is
available to train a model. We compare the ideal high-resource
scenario on English with results obtained on a smaller subset of the
same data and also run experiments on two small news corpora
covering low-resource languages, Croatian and Estonian. Two options
for headline generation in a multilingual low-resource scenario are
investigated: a pretrained multilingual encoder-decoder model and a
combination of two pretrained language models, one used as an
encoder and the other as a decoder, connected with a cross-attention
layer that needs to be trained from scratch. The results show that
the first approach outperforms the second one by a large margin. We
explore several data augmentation and pretraining strategies in
order to improve the performance of both models and show that while
we can drastically improve the second approach using these
strategies, they have little to no effect on the performance of the
pretrained encoder-decoder model. Finally, we propose two new
measures for evaluating the performance of the models besides the
classic ROUGE scores.},
  url   = {https://aclanthology.org/2022.lrec-1.381}
}

@InProceedings{zhou-portet-ringeval:2022:LREC,
  author   = {Zhou, Yongxin and Portet, François and Ringeval,
Fabien},
  title   = {Effectiveness of French Language Models on
Abstractive Dialogue Summarization Task},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3571--3581},
  abstract = {Pre-trained language models have established the
state-of-the-art on various natural language processing tasks,
including dialogue summarization, which allows the reader to quickly
access key information from long conversations in meetings,
interviews or phone calls. However, such dialogues are still
difficult to handle with current models because the spontaneity of
the language involves expressions that are rarely present in the
corpora used for pre-training the language models. Moreover, the

vast majority of the work accomplished in this field has been focused on English. In this work, we present a study on the summarization of spontaneous oral dialogues in French using several language specific pre-trained models: BARThez, and BelGPT-2, as well as multilingual pre-trained models: mBART, mBARThez, and mT5. Experiments were performed on the DECODA (Call Center) dialogue corpus whose task is to generate abstractive synopses from call center conversations between a caller and one or several agents depending on the situation. Results show that the BARThez models offer the best performance far above the previous state-of-the-art on DECODA. We further discuss the limits of such pre-trained models and the challenges that must be addressed for summarizing spontaneous dialogues.},
  url   = {https://aclanthology.org/2022.lrec-1.382}
}

@InProceedings{ferrs-saggion:2022:LREC,
  author   = {Ferrés, Daniel and Saggion, Horacio},
  title   = {ALEXSIS: A Dataset for Lexical Simplification in Spanish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3582--3594},
  abstract = {Lexical Simplification is the process of reducing the lexical complexity of a text by replacing difficult words with easier to read (or understand) expressions while preserving the original information and meaning. In this paper we introduce ALEXSIS, a new dataset for this task, and we use ALEXSIS to benchmark Lexical Simplification systems in Spanish. The paper describes the evaluation of three kind of approaches to Lexical Simplification, a thesaurus-based approach, a single transformers-based approach, and a combination of transformers. We also report state of the art results on a previous Lexical Simplification dataset for Spanish.},
  url   = {https://aclanthology.org/2022.lrec-1.383}
}

@InProceedings{mckinnon-rubino:2022:LREC,
  author   = {Mckinnon, Timothy and Rubino, Carl},
  title   = {The IARPA BETTER Program Abstract Task Four New Semantically Annotated Corpora from IARPA’s BETTER Program},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3595--3600},
  abstract = {IARPA’s Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program created multiple multilingual datasets to

spawn and evaluate cross-language information extraction and information retrieval research and development in zero-shot conditions. The first set of these resources for information extraction, the “Abstract” data will be released to the public at LREC 2022 in four languages to champion further information extraction work in this area. This paper presents the event and argument annotation in the Abstract Evaluation phase of BETTER, as well as the data collection, preparation, partitioning and mark-up of the datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.384}
}

@InProceedings{phan-nguyen-nguyen:2022:LREC,
  author   = {Phan, Uyen and Nguyen, Phuong N.V and Nguyen, Nhung},
  title   = {A Named Entity Recognition Corpus for Vietnamese Biomedical Texts to Support Tuberculosis Treatment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3601--3609},
  abstract = {Named Entity Recognition (NER) is an important task in information extraction. However, due to the lack of labelled corpora, biomedical NER has scarcely been studied in Vietnamese compared to English. To address this situation, we have constructed VietBioNER, a labelled NER corpus of Vietnamese academic biomedical text. The corpus focuses specifically on supporting tuberculosis surveillance, and was constructed by collecting scientific papers and grey literature related to tuberculosis symptoms and diagnostics. We manually annotated a small set of the collected documents with five categories of named entities: Organisation, Location, Date and Time, Symptom and Disease, and Diagnostic Procedure. Inter-annotator agreement ranges from 70.59\% and 95.89\% F-score according to entity category. In this paper, we make available two splits of the corpus, corresponding to traditional supervised learning and few-shot learning settings. We also provide baseline results for both of these settings, in addition to a dictionary-based approach, as a means to stimulate further research into Vietnamese biomedical NER. Although supervised methods produce results that are far superior to the other two approaches, the fact that even one-shot learning can outperform the dictionary-based method provides evidence that further research into few-shot learning on this text type would be worthwhile.},
  url   = {https://aclanthology.org/2022.lrec-1.385}
}

@InProceedings{mendezguzman-schlegel-batistanavarro:2022:LREC,
  author   = {Mendez Guzman, Erick and Schlegel, Viktor and Batista-Navarro, Riza},
  title   = {RaFoLa: A Rationale-Annotated Corpus for Detecting Indicators of Forced Labour},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3610--3625},
  abstract = {Forced labour is the most common type of modern slavery, and it is increasingly gaining the attention of the research and social community. Recent studies suggest that artificial intelligence (AI) holds immense potential for augmenting anti-slavery action. However, AI tools need to be developed transparently in cooperation with different stakeholders. Such tools are contingent on the availability and access to domain-specific data, which are scarce due to the near-invisible nature of forced labour. To the best of our knowledge, this paper presents the first openly accessible English corpus annotated for multi-class and multi-label forced labour detection. The corpus consists of 989 news articles retrieved from specialised data sources and annotated according to risk indicators defined by the International Labour Organization (ILO). Each news article was annotated for two aspects: (1) indicators of forced labour as classification labels and (2) snippets of the text that justify labelling decisions. We hope that our data set can help promote research on explainability for multi-class and multi-label text classification. In this work, we explain our process for collecting the data underpinning the proposed corpus, describe our annotation guidelines and present some statistical analysis of its content. Finally, we summarise the results of baseline experiments based on different variants of the Bidirectional Encoder Representation from Transformer (BERT) model.},
  url   = {https://aclanthology.org/2022.lrec-1.386}
}

@InProceedings{jarrar-khalilia-ghanem:2022:LREC,
  author   = {Jarrar, Mustafa and Khalilia, Mohammed and Ghanem, Sana},
  title   = {Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3626--3636},
  abstract = {This paper presents Wojood, a corpus for Arabic nested Named Entity Recognition (NER). Nested entities occur when one entity mention is embedded inside another entity mention. Wojood consists of about 550K Modern Standard Arabic (MSA) and dialect tokens that are manually annotated with 21 entity types including person, organization, location, event and date. More importantly, the corpus is annotated with nested entities instead of the more common flat annotations. The data contains about 75K entities and

22.5\% of which are nested. The inter-annotator evaluation of the corpus demonstrated a strong agreement with Cohen's Kappa of 0.979 and an F1-score of 0.976. To validate our data, we used the corpus to train a nested NER model based on multi-task learning using the pre-trained AraBERT (Arabic BERT). The model achieved an overall micro F1-score of 0.884. Our corpus, the annotation guidelines, the source code and the pre-trained model are publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.387}
}

@InProceedings{raithel-EtAl:2022:LREC,
  author   = {Raithel, Lisa and Thomas, Philippe and Roller, Roland and Sapina, Oliver and Möller, Sebastian and Zweigenbaum, Pierre},
  title   = {Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3637--3649},
  abstract = {In this work, we present the first corpus for German Adverse Drug Reaction (ADR) detection in patient-generated content. The data consists of 4,169 binary annotated documents from a German patient forum, where users talk about health issues and get advice from medical doctors. As is common in social media data in this domain, the class labels of the corpus are very imbalanced. This and a high topic imbalance make it a very challenging dataset, since often, the same symptom can have several causes and is not always related to a medication intake. We aim to encourage further multi-lingual efforts in the domain of ADR detection and provide preliminary experiments for binary classification using different methods of zero- and few-shot learning based on a multi-lingual model. When fine-tuning XLM-RoBERTa first on English patient forum data and then on the new German data, we achieve an F1-score of 37.52 for the positive class. We make the dataset and models publicly available for the community.},
  url   = {https://aclanthology.org/2022.lrec-1.388}
}

@InProceedings{borchert-EtAl:2022:LREC,
  author   = {Borchert, Florian and Lohr, Christina and Modersohn, Luise and Witt, Jonas and Langer, Thomas and Follmann, Markus and Gietzelt, Matthias and Arnrich, Bert and Hahn, Udo and Schapranow, Matthieu-P.},
  title   = {GGPONC 2.0 - The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3650--3660},
abstract = {Despite remarkable advances in the development of language resources over the recent years, there is still a shortage of annotated, publicly available corpora covering (German) medical language. With the initial release of the German Guideline Program in Oncology NLP Corpus (GGPONC), we have demonstrated how such corpora can be built upon clinical guidelines, a widely available resource in many natural languages with a reasonable coverage of medical terminology. In this work, we describe a major new release for GGPONC. The corpus has been substantially extended in size and re-annotated with a new annotation scheme based on SNOMED CT top level hierarchies, reaching high inter-annotator agreement (γ=.94). Moreover, we annotated elliptical coordinated noun phrases and their resolutions, a common language phenomenon in (not only German) scientific documents. We also trained BERT-based named entity recognition models on this new data set, which achieve high performance on short, coarse-grained entity spans (F1=.89), while the rate of boundary errors increases for long entity spans. GGPONC is freely available through a data use agreement. The trained named entity recognition models, as well as the detailed annotation guide, are also made publicly available.},
url = {https://aclanthology.org/2022.lrec-1.389}
}

@InProceedings{avram-EtAl:2022:LREC,
author = {Avram, Andrei-Marius and Catrina, Darius and Cercel, Dumitru-Clementin and Dascalu, Mihai and Rebedea, Traian and Pais, Vasile and Tufis, Dan},
title = {Distilling the Knowledge of Romanian BERTs Using Multiple Teachers},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {374--384},
abstract = {Running large-scale pre-trained language models in computationally constrained environments remains a challenging problem yet to be addressed, while transfer learning from these models has become prevalent in Natural Language Processing tasks. Several solutions, including knowledge distillation, network quantization, or network pruning have been previously proposed; however, these approaches focus mostly on the English language, thus widening the gap when considering low-resource languages. In this work, we introduce three light and fast versions of distilled BERT models for the Romanian language: Distil-BERT-base-ro, Distil-RoBERT-base, and DistilMulti-BERT-base-ro. The first two models resulted from the individual distillation of knowledge from two base versions of Romanian BERTs available in literature, while the last one was obtained by distilling their ensemble. To our knowledge, this is the first attempt to create publicly available Romanian

distilled BERT models, which were thoroughly evaluated on five tasks: part-of-speech tagging, named entity recognition, sentiment analysis, semantic textual similarity, and dialect identification. Our experimental results argue that the three distilled models offer performance comparable to their teachers, while being twice as fast on a GPU and ~35\% smaller. In addition, we further test the similarity between the predictions of our students versus their teachers by measuring their label and probability loyalty, together with regression loyalty - a new metric introduced in this work.},
  url   = {https://aclanthology.org/2022.lrec-1.39}
}

@InProceedings{zotova-cuadros-rigau:2022:LREC,
  author   = {Zotova, Elena and Cuadros, Montse and Rigau, German},
  title   = {ClinIDMap: Towards a Clinical IDs Mapping for Data Interoperability},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3661--3669},
  abstract = {This paper presents ClinIDMap, a tool for mapping identifiers between clinical ontologies and lexical resources. ClinIDMap interlinks identifiers from UMLS, SMOMED-CT, ICD-10 and the corresponding Wikipedia articles for concepts from the UMLS Metathesaurus. Our main goal is to provide semantic interoperability across the clinical concepts from various knowledge bases. As a side effect, the mapping enriches already annotated corpora in multiple languages with new labels. For instance, spans manually annotated with IDs from UMLS can be annotated with Semantic Types and Groups, and its corresponding SNOMED CT and ICD-10 IDs. We also experiment with sequence labelling models for detecting Diagnosis and Procedures concepts and for detecting UMLS Semantic Groups trained on Spanish, English, and bilingual corpora obtained with the new mapping procedure. The ClinIDMap tool is publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.390}
}

@InProceedings{ceausu-nisioi:2022:LREC,
  author   = {Ceausu, Corina and Nisioi, Sergiu},
  title   = {Identifying Draft Bills Impacting Existing Legislation: a Case Study on Romanian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3670--3674},
  abstract = {In our paper, we present a novel corpus of historical legal documents on the Romanian public procurement legislation and

an annotated subset of draft bills that have been screened by legal experts and identified as impacting past public procurement legislation. Using the manual annotations provided by the experts, we attempt to automatically identify future draft bills that have the potential to impact existing policies on public procurement.},
  url   = {https://aclanthology.org/2022.lrec-1.391}
}

@InProceedings{hudson-almoubayed:2022:LREC,
  author   = {Hudson, George and Al Moubayed, Noura},
  title   = {MuLD: The Multitask Long Document Benchmark},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3675--3685},
  abstract = {The impressive progress in NLP techniques has been driven by the development of multi-task benchmarks such as GLUE and SuperGLUE. While these benchmarks focus on tasks for one or two input sentences, there has been exciting work in designing efficient techniques for processing much longer inputs. In this paper, we present MuLD: a new long document benchmark consisting of only documents over 10,000 tokens. By modifying existing NLP tasks, we create a diverse benchmark which requires models to successfully model long-term dependencies in the text. We evaluate how existing models perform, and find that our benchmark is much more challenging than their 'short document' equivalents. Furthermore, by evaluating both regular and efficient transformers, we show that models with increased context length are better able to solve the tasks presented, suggesting that future improvements in these models are vital for solving similar long document problems. We release the data and code for baselines to encourage further research on efficient NLP models.},
  url   = {https://aclanthology.org/2022.lrec-1.392}
}

@InProceedings{datta-EtAl:2022:LREC,
  author   = {Datta, Surabhi and Lam, Hio Cheng and Pajouhi, Atieh and Mogalla, Sunitha and Roberts, Kirk},
  title   = {A Cross-document Coreference Dataset for Longitudinal Tracking across Radiology Reports},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3686--3695},
  abstract = {This paper proposes a new cross-document coreference resolution (CDCR) dataset for identifying co-referring radiological findings and medical devices across a patient's radiology reports. Our annotated corpus contains 5872 mentions (findings and devices)

spanning 638 MIMIC-III radiology reports across 60 patients, covering multiple imaging modalities and anatomies. There are a total of 2292 mention chains. We describe the annotation process in detail, highlighting the complexities involved in creating a sizable and realistic dataset for radiology CDCR. We apply two baseline methods—string matching and transformer language models (BERT)—to identify cross-report coreferences. Our results indicate the requirement of further model development targeting better understanding of domain language and context to address this challenging and unexplored task. This dataset can serve as a resource to develop more advanced natural language processing CDCR methods in the future. This is one of the first attempts focusing on CDCR in the clinical domain and holds potential in benefiting physicians and clinical research through long-term tracking of radiology findings.},
  url   = {https://aclanthology.org/2022.lrec-1.393}
}

@InProceedings{khaldi-EtAl:2022:LREC,
  author   = {Khaldi, Hadjer and Benamara, Farah and Pradel, Camille and Sigel, Grégoire and Aussenac-Gilles, Nathalie},
  title   = {How's Business Going Worldwide ? A Multilingual Annotated Corpus for Business Relation Extraction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3696--3705},
  abstract = {The business world has changed due to the 21st century economy, where borders have melted and trades became free. Nowadays,competition is no longer only at the local market level but also at the global level. In this context, the World Wide Web has become a major source of information for companies and professionals to keep track of their complex, rapidly changing, and competitive business environment. A lot of effort is nonetheless needed to collect and analyze this information due to information overload problem and the huge number of web pages to process and analyze. In this paper, we propose the BizRel resource, the first multilingual (French,English, Spanish, and Chinese) dataset for automatic extraction of binary business relations involving organizations from the web.This dataset is used to train several monolingual and cross-lingual deep learning models to detect these relations in texts. Our results are encouraging, demonstrating the effectiveness of such a resource for both research and business communities. In particular, we believe multilingual business relation extraction systems are crucial tools for decision makers to identify links between specific market stakeholders and build business networks which enable to anticipate changes and discover new threats or opportunities. Our work is therefore an important direction toward such tools.},
  url   = {https://aclanthology.org/2022.lrec-1.394}
}

@InProceedings{rahimi-surdeanu:2022:LREC,
  author   = {Rahimi, Mahdi and Surdeanu, Mihai},
  title   = {Do Transformer Networks Improve the Discovery of Rules from Text?},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3706--3714},
  abstract = {With their Discovery of Inference Rules from Text (DIRT) algorithm, Lin and Pantel (2001) made a seminal contribution to the field of rule acquisition from text, by adapting the distributional hypothesis of Harris (1954) to rules that model binary relations such as X treat Y. DIRT’s relevance is renewed in today’s neural era given the recent focus on interpretability in the field of natural language processing. We propose a novel take on the DIRT algorithm, where we implement the distributional hypothesis using the contextualized embeddings provided by BERT, a transformer-network-based language model (Vaswani et al. 2017; Devlin et al. 2018). In particular, we change the similarity measure between pairs of slots (i.e., the set of words matched by a rule) from the original formula that relies on lexical items to a formula computed using contextualized embeddings. We empirically demonstrate that this new similarity method yields a better implementation of the distributional hypothesis, and this, in turn, yields rules that outperform the original algorithm in the question answering-based evaluation proposed by Lin and Pantel (2001).},
  url   = {https://aclanthology.org/2022.lrec-1.395}
}

@InProceedings{litvak-EtAl:2022:LREC,
  author   = {Litvak, Marina and Vanetik, Natalia and Liebeskind, Chaya and Hmdia, Omar and Madeghem, Rizek Abu},
  title   = {Offensive language detection in Hebrew: can other languages help?},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3715--3723},
  abstract = {Unfortunately, offensive language in social media is a common phenomenon nowadays. It harms many people and vulnerable groups. Therefore, automated detection of offensive language is in high demand and it is a serious challenge in multilingual domains. Various machine learning approaches combined with natural language techniques have been applied for this task lately. This paper contributes to this area from several aspects: (1) it introduces a new dataset of annotated Facebook comments in Hebrew; (2) it describes a case study with multiple supervised models and text representations for a task of offensive language detection in three

languages, including two Semitic (Hebrew and Arabic) languages; (3) it reports evaluation results of cross-lingual and multilingual learning for detection of offensive content in Semitic languages; and (4) it discusses the limitations of these settings.},
  url   = {https://aclanthology.org/2022.lrec-1.396}
}

@InProceedings{cheng-EtAl:2022:LREC2,
  author   = {Cheng, Fei and Yada, Shuntaro and Tanaka, Ribeka and ARAMAKI, Eiji and Kurohashi, Sadao},
  title   = {JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3724--3731},
  abstract = {In the field of Japanese medical information extraction, few analyzing tools are available and relation extraction is still an under-explored topic. In this paper, we first propose a novel relation annotation schema for investigating the medical and temporal relations between medical entities in Japanese medical reports. We experiment with the practical annotation scenarios by separately annotating two different types of reports. We design a pipeline system with three components for recognizing medical entities, classifying entity modalities, and extracting relations. The empirical results show accurate analyzing performance and suggest the satisfactory annotation quality, the superiority of the latest contextual embedding models. and the feasible annotation strategy for high-accuracy demand.},
  url   = {https://aclanthology.org/2022.lrec-1.397}
}

@InProceedings{strobl-trabelsi-zaane:2022:LREC,
  author   = {Strobl, Michael and Trabelsi, Amine and Zaïane, Osmar},
  title   = {Enhanced Entity Annotations for Multilingual Corpora},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3732--3740},
  abstract = {Modern approaches in Natural Language Processing (NLP) require, ideally, large amounts of labelled data for model training. However, new language resources, for example, for Named Entity Recognition (NER), Co-reference Resolution (CR), Entity Linking (EL) and Relation Extraction (RE), naming a few of the most popular tasks in NLP, have always been challenging to create since manual text annotations can be very time-consuming to acquire. While

there may be an acceptable amount of labelled data available for some of these tasks in one language, there may be a lack of datasets in another. WEXEA is a tool to exhaustively annotate entities in the English Wikipedia. Guidelines for editors of Wikipedia articles result, on the one hand, in only a few annotations through hyperlinks, but on the other hand, make it easier to exhaustively annotate the rest of these articles with entities than starting from scratch. We propose the following main improvements to WEXEA: Creating multi-lingual corpora, improved entity annotations using a proven NER system, annotating dates and times. A brief evaluation of the annotation quality of WEXEA is added.},
  url   = {https://aclanthology.org/2022.lrec-1.398}
}

@InProceedings{menya-EtAl:2022:LREC,
  author   = {Menya, Edmond and Roche, Mathieu and Interdonato, Roberto and Owuor, Dickson},
  title   = {Enriching Epidemiological Thematic Features For Disease Surveillance Corpora Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3741--3750},
  abstract = {We present EpidBioBERT, a biosurveillance epidemiological document tagger for disease surveillance over PADI-Web system. Our model is trained on PADI-Web corpus which contains news articles on Animal Diseases Outbreak extracted from the web. We train a classifier to discriminate between relevant and irrelevant documents based on their epidemiological thematic feature content in preparation for further epidemiology information extraction. Our approach proposes a new way to perform epidemiological document classification by enriching epidemiological thematic features namely disease, host, location and date, which are used as inputs to our epidemiological document classifier. We adopt a pre-trained biomedical language model with a novel fine tuning approach that enriches these epidemiological thematic features. We find these thematic features rich enough to improve epidemiological document classification over a smaller data set than initially used in PADI-Web classifier. This improves the classifiers ability to avoid false positive alerts on disease surveillance systems. To further understand information encoded in EpidBioBERT, we experiment the impact of each epidemiology thematic feature on the classifier under ablation studies. We compare our biomedical pre-trained approach with a general language model based model finding that thematic feature embeddings pre-trained on general English documents are not rich enough for epidemiology classification task. Our model achieves an F1-score of 95.5\% over an unseen test set, with an improvement of +5.5 points on F1-Score on the PADI-Web classifier with nearly half the training data set.},
  url   = {https://aclanthology.org/2022.lrec-1.399}
}

@InProceedings{colman-EtAl:2022:LREC,
  author   = {Colman, Toon and Fonteyne, Margot and Daems, Joke and Dirix, Nicolas and Macken, Lieve},
  title   = {GECO-MT: The Ghent Eye-tracking Corpus of Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {29--38},
  abstract = {In the present paper, we describe a large corpus of eye movement data, collected during natural reading of a human translation and a machine translation of a full novel. This data set, called GECO-MT (Ghent Eye tracking Corpus of Machine Translation) expands upon an earlier corpus called GECO (Ghent Eye-tracking Corpus) by Cop et al. (2017). The eye movement data in GECO-MT will be used in future research to investigate the effect of machine translation on the reading process and the effects of various error types on reading. In this article, we describe in detail the materials and data collection procedure of GECO-MT. Extensive information on the language proficiency of our participants is given, as well as a comparison with the participants of the original GECO. We investigate the distribution of a selection of important eye movement variables and explore the possibilities for future analyses of the data. GECO-MT is freely available at https://www.lt3.ugent.be/resources/geco-mt.},
  url   = {https://aclanthology.org/2022.lrec-1.4}
}

@InProceedings{matsunaga-EtAl:2022:LREC,
  author   = {Matsunaga, Yuta and Saeki, Takaaki and Takamichi, Shinnosuke and Saruwatari, Hiroshi},
  title   = {Personalized Filled-pause Generation with Group-wise Prediction Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {385--392},
  abstract = {In this paper, we propose a method to generate personalized filled pauses (FPs) with group-wise prediction models. Compared with fluent text generation, disfluent text generation has not been widely explored. To generate more human-like texts, we addressed disfluent text generation. The usage of disfluency, such as FPs, rephrases, and word fragments, differs from speaker to speaker, and thus, the generation of personalized FPs is required. However, it is difficult to predict them because of the sparsity of position and the frequency difference between more and less frequently used FPs. Moreover, it is sometimes difficult to adapt FP

prediction models to each speaker because of the large variation of the tendency within each speaker. To address these issues, we propose a method to build group-dependent prediction models by grouping speakers on the basis of their tendency to use FPs. This method does not require a large amount of data and time to train each speaker model. We further introduce a loss function and a word embedding model suitable for FP prediction. Our experimental results demonstrate that group-dependent models can predict FPs with higher scores than a non-personalized one and the introduced loss function and word embedding model improve the prediction performance.},
  url   = {https://aclanthology.org/2022.lrec-1.40}
}

@InProceedings{degibertbonet-EtAl:2022:LREC2,
  author   = {de Gibert Bonet, Ona and García Pablos, Aitor and Cuadros, Montse and Melero, Maite},
  title   = {Spanish Datasets for Sensitive Entity Detection in the Legal Domain},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3751--3760},
  abstract = {The de-identification of sensible data, also known as automatic textual anonymisation, is essential for data sharing and reuse, both for research and commercial purposes. The first step for data anonymisation is the detection of sensible entities. In this work, we present four new datasets for named entity detection in Spanish in the legal domain. These datasets have been generated in the framework of the MAPA project, three smaller datasets have been manually annotated and one large dataset has been automatically annotated, with an estimated error rate of around 14\%. In order to assess the quality of the generated datasets, we have used them to fine-tune a battery of entity-detection models, using as foundation different pre-trained language models: one multilingual, two general-domain monolingual and one in-domain monolingual. We compare the results obtained, which validate the datasets as a valuable resource to fine-tune models for the task of named entity detection. We further explore the proposed methodology by applying it to a real use case scenario.},
  url   = {https://aclanthology.org/2022.lrec-1.400}
}

@InProceedings{bhattarai-granmo-jiao:2022:LREC1,
  author   = {Bhattarai, Bimal and Granmo, Ole-Christoffer and Jiao, Lei},
  title   = {ConvTextTM: An Explainable Convolutional Tsetlin Machine Framework for Text Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3761--3770},
abstract = {Recent advancements in natural language processing (NLP) have reshaped the industry, with powerful language models such as GPT-3 achieving superhuman performance on various tasks. However, the increasing complexity of such models turns them into "black boxes'', creating uncertainty about their internal operation and decision-making. Tsetlin Machine (TM) employs human-interpretable conjunctive clauses in propositional logic to solve complex pattern recognition problems and has demonstrated competitive performance in various NLP tasks. In this paper, we propose ConvTextTM, a novel convolutional TM architecture for text classification. While legacy TM solutions treat the whole text as a corpus-specific set-of-words (SOW), ConvTextTM breaks down the text into a sequence of text fragments. The convolution over the text fragments opens up for local position-aware analysis. Further, ConvTextTM eliminates the dependency on a corpus-specific vocabulary. Instead, it employs a generic SOW formed by the tokenization scheme of the Bidirectional Encoder Representations from Transformers (BERT). The convolution binds together the tokens, allowing ConvTextTM to address the out-of-vocabulary problem as well as spelling errors. We investigate the local explainability of our proposed method using clause-based features. Extensive experiments are conducted on seven datasets, to demonstrate that the accuracy of ConvTextTM is either superior or comparable to state-of-the-art baselines.},
url = {https://aclanthology.org/2022.lrec-1.401}
}

@InProceedings{beloucif-EtAl:2022:LREC,
author = {Beloucif, Meriem and Yimam, Seid Muhie and Stahlhacke, Steffen and Biemann, Chris},
title = {Elvis vs. M. Jackson: Who has More Albums? Classification and Identification of Elements in Comparative Questions},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3771--3779},
abstract = {Comparative Question Answering (cQA) is the task of providing concrete and accurate responses to queries such as: ``Is Lyft cheaper than a regular taxi?'' or ``What makes a mortgage different from a regular loan?". In this paper, we propose two new open-domain real-world datasets for identifying and labeling comparative questions. While the first dataset contains instances of English questions labeled as comparative vs. non-comparative, the second dataset provides additional labels including the objects and the aspects of comparison. We conduct several experiments that evaluate the soundness of our datasets. The evaluation of our datasets using various classifiers show promising results that reach close-to-human results on a binary classification task with a neural

model using ALBERT embeddings. When approaching the unsupervised sequence labeling task, some headroom remains.},
  url   = {https://aclanthology.org/2022.lrec-1.402}
}

@InProceedings{yeh-lavergne-zweigenbaum:2022:LREC,
  author   = {Yeh, Hui-Syuan and Lavergne, Thomas and Zweigenbaum, Pierre},
  title   = {Decorate the Examples: A Simple Method of Prompt Design for Biomedical Relation Extraction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3780--3787},
  abstract = {Relation extraction is a core problem for natural language processing in the biomedical domain. Recent research on relation extraction showed that prompt-based learning improves the performance on both fine-tuning on full training set and few-shot training. However, less effort has been made on domain-specific tasks where good prompt design can be even harder. In this paper, we investigate prompting for biomedical relation extraction, with experiments on the ChemProt dataset. We present a simple yet effective method to systematically generate comprehensive prompts that reformulate the relation extraction task as a cloze-test task under a simple prompt formulation. In particular, we experiment with different ranking scores for prompt selection. With BioMed-RoBERTa-base, our results show that prompting-based fine-tuning obtains gains by 14.21 F1 over its regular fine-tuning baseline, and 1.14 F1 over SciFive-Large, the current state-of-the-art on ChemProt. Besides, we find prompt-based learning requires fewer training examples to make reasonable predictions. The results demonstrate the potential of our methods in such a domain-specific relation extraction task.},
  url   = {https://aclanthology.org/2022.lrec-1.403}
}

@InProceedings{ivanova-vanerp-kirrane:2022:LREC,
  author   = {Ivanova, Rositsa and van Erp, Marieke and Kirrane, Sabrina},
  title   = {Comparing Annotated Datasets for Named Entity Recognition in English Literature},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3788--3797},
  abstract = {The growing interest in named entity recognition (NER) in various domains has led to the creation of different benchmark datasets, often with slightly different annotation

guidelines. To better understand the different NER benchmark datasets for the domain of English literature and their impact on the evaluation of NER tools, we analyse two existing annotated datasets and create two additional gold standard datasets. Following on from this, we evaluate the performance of two NER tools, one domain-specific and one general-purpose NER tool, using the four gold standards, and analyse the sources for the differences in the measured performance. Our results show that the performance of the two tools varies significantly depending on the gold standard used for the individual evaluations.},
  url   = {https://aclanthology.org/2022.lrec-1.404}
}

@InProceedings{sakketou-EtAl:2022:LREC2,
  author   = {Sakketou, Flora and Lahnala, Allison and Vogel, Liane and Flek, Lucie},
  title   = {Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3798--3808},
  abstract = {There is an increasing need for the ability to model fine-grained opinion shifts of social media users, as concerns about the potential polarizing social effects increase. However, the lack of publicly available datasets that are suitable for the task presents a major challenge. In this paper, we introduce an innovative annotated dataset for modeling subtle opinion fluctuations and detecting fine-grained stances. The dataset includes a sufficient amount of stance polarity and intensity labels per user over time and within entire conversational threads, thus making subtle opinion fluctuations detectable both in long term and in short term. All posts are annotated by non-experts and a significant portion of the data is also annotated by experts. We provide a strategy for recruiting suitable non-experts. Our analysis of the inter-annotator agreements shows that the resulting annotations obtained from the majority vote of the non-experts are of comparable quality to the annotations of the experts. We provide analyses of the stance evolution in short term and long term levels, a comparison of language usage between users with vacillating and resolute attitudes, and fine-grained stance detection baselines.},
  url   = {https://aclanthology.org/2022.lrec-1.405}
}

@InProceedings{stranisci-EtAl:2022:LREC,
  author   = {Stranisci, Marco Antonio and Frenda, Simona and Ceccaldi, Eleonora and Basile, Valerio and Damiano, Rossana and Patti, Viviana},
  title   = {APPReddit: a Corpus of Reddit Posts Annotated for Appraisal},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3809--3818},
  abstract = {Despite the large number of computational resources for emotion recognition, there is a lack of data sets relying on appraisal models. According to Appraisal theories, emotions are the outcome of a multi-dimensional evaluation of events. In this paper, we present APPReddit, the first corpus of non-experimental data annotated according to this theory. After describing its development, we compare our resource with enISEAR, a corpus of events created in an experimental setting and annotated for appraisal. Results show that the two corpora can be mapped notwithstanding different typologies of data and annotations schemes. A SVM model trained on APPReddit predicts four appraisal dimensions without significant loss. Merging both corpora in a single training set increases the prediction of 3 out of 4 dimensions. Such findings pave the way to a better performing classification model for appraisal prediction.},
  url   = {https://aclanthology.org/2022.lrec-1.406}
}

@InProceedings{machado-pardo:2022:LREC,
  author   = {Machado, Mateus and Pardo, Thiago Alexandre Salgueiro},
  title   = {Evaluating Methods for Extraction of Aspect Terms in Opinion Texts in Portuguese - the Challenges of Implicit Aspects},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3819--3828},
  abstract = {One of the challenges of aspect-based sentiment analysis is the implicit mention of aspects. These are more difficult to identify and may require world knowledge to do so. In this work, we evaluate frequency-based, hybrid, and machine learning methods, including the use of the pre-trained BERT language model, in the task of extracting aspect terms in opinionated texts in Portuguese, emphasizing the analysis of implicit aspects. Besides the comparative evaluation of methods, the differential of this work lies in the analysis's novelty using a typology of implicit aspects that shows the knowledge needed to identify each implicit aspect term, thus allowing a mapping of the strengths and weaknesses of each method.},
  url   = {https://aclanthology.org/2022.lrec-1.407}
}

@InProceedings{cambria-EtAl:2022:LREC,
  author   = {Cambria, Erik and Liu, Qian and Decherchi, Sergio and Xing, Frank and Kwok, Kenneth},

  title   = {SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3829--3839},
  abstract = {In recent years, AI research has demonstrated enormous potential for the benefit of humanity and society. While often better than its human counterparts in classification and pattern recognition tasks, however, AI still struggles with complex tasks that require commonsense reasoning such as natural language understanding. In this context, the key limitations of current AI models are: dependency, reproducibility, trustworthiness, interpretability, and explainability. In this work, we propose a commonsense-based neurosymbolic framework that aims to overcome these issues in the context of sentiment analysis. In particular, we employ unsupervised and reproducible subsymbolic techniques such as auto-regressive language models and kernel methods to build trustworthy symbolic representations that convert natural language to a sort of protolanguage and, hence, extract polarity from text in a completely interpretable and explainable manner.},
  url   = {https://aclanthology.org/2022.lrec-1.408}
}

@InProceedings{zariquiey-EtAl:2022:LREC,
  author   = {Zariquiey, Roberto and Alvarado, Claudia and Echevarría, Ximena and Gomez, Luisa and Gonzales, Rosa and Illescas, Mariana and Oporto, Sabina and Blum, Frederic and Oncevay, Arturo and Vera, Javier},
  title   = {Building an Endangered Language Resource in the Classroom: Universal Dependencies for Kakataibo},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3840--3851},
  abstract = {In this paper, we launch a new Universal Dependencies treebank for an endangered language from Amazonia: Kakataibo, a Panoan language spoken in Peru. We first discuss the collaborative methodology implemented, which proved effective to create a treebank in the context of a Computational Linguistic course for undergraduates. Then, we describe the general details of the treebank and the language-specific considerations implemented for the proposed annotation. We finally conduct some experiments on part-of-speech tagging and syntactic dependency parsing. We focus on monolingual and transfer learning settings, where we study the impact of a Shipibo-Konibo treebank, another Panoan language resource.},
  url   = {https://aclanthology.org/2022.lrec-1.409}

}

@InProceedings{sheikh-vincent-illina:2022:LREC,
  author   = {Sheikh, Imran and Vincent, Emmanuel and Illina, Irina},
  title   = {Transformer versus LSTM Language Models trained on Uncertain ASR Hypotheses in Limited Data Scenarios},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {393--399},
  abstract = {In several ASR use cases, training and adaptation of domain-specific LMs can only rely on a small amount of manually verified text transcriptions and sometimes a limited amount of in-domain speech. Training of LSTM LMs in such limited data scenarios can benefit from alternate uncertain ASR hypotheses, as observed in our recent work. In this paper, we propose a method to train Transformer LMs on ASR confusion networks. We evaluate whether these self-attention based LMs are better at exploiting alternate ASR hypotheses as compared to LSTM LMs. Evaluation results show that Transformer LMs achieve 3-6\% relative reduction in perplexity on the AMI scenario meetings but perform similar to LSTM LMs on the smaller Verbmobil conversational corpus. Evaluation on ASR N-best rescoring shows that LSTM and Transformer LMs trained on ASR confusion networks do not bring significant WER reductions. However, a qualitative analysis reveals that they are better at predicting less frequent words.},
  url   = {https://aclanthology.org/2022.lrec-1.41}
}

@InProceedings{kummervold-wetjen-delarosa:2022:LREC,
  author   = {Kummervold, Per and Wetjen, Freddy and de la Rosa, Javier},
  title   = {The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3852--3860},
  abstract = {Norwegian has been one of many languages lacking sufficient available text to train quality language models. In an attempt to bridge this gap, we introduce the Norwegian Colossal Corpus (NCC), which comprises 49GB of clean Norwegian textual data containing over 7B words. The NCC is composed of different and varied sources, ranging from books and newspapers to government documents and public reports, showcasing the various uses of the Norwegian language in society. The corpus contains mainly Norwegian Bokmål and Norwegian Nynorsk. Each document in the corpus is tagged

with metadata that enables the creation of sub-corpora for specific needs. Its structure makes it easy to combine with large web archives that for licensing reasons could not be distributed together with the NCC. By releasing this corpus openly to the public, we hope to foster the creation of both better Norwegian language models and multilingual language models with support for Norwegian.},
  url   = {https://aclanthology.org/2022.lrec-1.410}
}

@InProceedings{lugli-EtAl:2022:LREC,
  author   = {Lugli, Ligeia and Martinc, Matej and Pelicon, Andraž and Pollak, Senja},
  title   = {Embeddings models for Buddhist Sanskrit},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3861--3871},
  abstract = {The paper presents novel resources and experiments for Buddhist Sanskrit, broadly defined here including all the varieties of Sanskrit in which Buddhist texts have been transmitted. We release a novel corpus of Buddhist texts, a novel corpus of general Sanskrit and word similarity and word analogy datasets for intrinsic evaluation of Buddhist Sanskrit embeddings models. We compare the performance of word2vec and fastText static embeddings models, with default and optimized parameter settings, as well as contextual models BERT and GPT-2, with different training regimes (including a transfer learning approach using the general Sanskrit corpus) and different embeddings construction regimes (given the encoder layers). The results show that for semantic similarity the fastText embeddings yield the best results, while for word analogy tasks BERT embeddings work the best. We also show that for contextual models the optimal layer combination for embedding construction is task dependant, and that pretraining the contextual embeddings models on a reference corpus of general Sanskrit is beneficial, which is a promising finding for future development of embeddings for less-resourced languages and domains.},
  url   = {https://aclanthology.org/2022.lrec-1.411}
}

@InProceedings{cotosolano-EtAl:2022:LREC,
  author   = {Coto-Solano, Rolando and Nicholas, Sally Akevai and Datta, Samiha and Quint, Victoria and Wills, Piripi and Powell, Emma Ngakuravaru and Koka'ua, Liam and Tanveer, Syed and Feldman, Isaac},
  title   = {Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3872--3882},
abstract = {This paper describes the process of data processing and training of an automatic speech recognition (ASR) system for Cook Islands Māori (CIM), an Indigenous language spoken by approximately 22,000 people in the South Pacific. We transcribed four hours of speech from adults and elderly speakers of the language and prepared two experiments. First, we trained three ASR systems: one statistical, Kaldi; and two based on Deep Learning, DeepSpeech and XLSR-Wav2Vec2. Wav2Vec2 tied with Kaldi for lowest character error rate (CER=6±1) and was slightly behind in word error rate (WER=23±2 versus WER=18±2 for Kaldi). This provides evidence that Deep Learning ASR systems are reaching the performance of statistical methods on small datasets, and that they can work effectively with extremely low-resource Indigenous languages like CIM. In the second experiment we used Wav2Vec2 to train models with held-out speakers. While the performance decreased (CER=15±7, WER=46±16), the system still showed considerable learning. We intend to use ASR to accelerate the documentation of CIM, using newly transcribed texts to improve the ASR and also generate teaching and language revitalization materials. The trained model is available under a license based on the Kaitiakitanga License, which provides for non-commercial use while retaining control of the model by the Indigenous community.},
url = {https://aclanthology.org/2022.lrec-1.412}
}

@InProceedings{wiedemann-EtAl:2022:LREC,
author = {Wiedemann, Gregor and Dollbaum, Jan Matti and Haunss, Sebastian and Daphi, Priska and Meier, Larissa Daria},
title = {A Generalized Approach to Protest Event Detection in German Local News},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {3883--3891},
abstract = {Protest events provide information about social and political conflicts, the state of social cohesion and democratic conflict management, as well as the state of civil society in general. Social scientists are therefore interested in the systematic observation of protest events. With this paper, we release the first German language resource of protest event related article excerpts published in local news outlets. We use this dataset to train and evaluate transformer-based text classifiers to automatically detect relevant newspaper articles. Our best approach reaches a binary F1-score of 93.3 \%, which is a promising result for our goal to support political science research. However, in a second experiment, we show that our model does not generalize equally well when applied to data from time periods and localities other than our training sample. To make protest event detection more

robust, we test two ways of alternative preprocessing. First, we find that letting the classifier concentrate on sentences around protest keywords improves the F1-score for out-of-sample data up to +4 percentage points. Second, against our initial intuition, masking of named entities during preprocessing does not improve the generalization in terms of F1-scores. However, it leads to a significantly improved recall of the models.},
  url   = {https://aclanthology.org/2022.lrec-1.413}
}

@InProceedings{gnehm-bhlmann-clematide:2022:LREC,
  author   = {Gnehm, Ann-Sophie and Bühlmann, Eva and Clematide, Simon},
  title   = {Evaluation of Transfer Learning and Domain Adaptation for Analyzing German-Speaking Job Advertisements},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3892--3901},
  abstract = {This paper presents text mining approaches on German-speaking job advertisements to enable social science research on the development of the labour market over the last 30 years. In order to build text mining applications providing information about profession and main task of a job, as well as experience and ICT skills needed, we experiment with transfer learning and domain adaptation. Our main contribution consists in building language models which are adapted to the domain of job advertisements, and their assessment on a broad range of machine learning problems. Our findings show the large value of domain adaptation in several respects. First, it boosts the performance of fine-tuned task-specific models consistently over all evaluation experiments. Second, it helps to mitigate rapid data shift over time in our special domain, and enhances the ability to learn from small updates with new, labeled task data. Third, domain-adaptation of language models is efficient: With continued in-domain pre-training we are able to outperform general-domain language models pre-trained on ten times more data. We share our domain-adapted language models and data with the research community.},
  url   = {https://aclanthology.org/2022.lrec-1.414}
}

@InProceedings{perezalmendros-espinosaanke-schockaert:2022:LREC,
  author   = {Perez Almendros, Carla and Espinosa Anke, Luis and Schockaert, Steven},
  title   = {Pre-Training Language Models for Identifying Patronizing and Condescending Language: An Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {3902--3911},
  abstract = {Patronizing and Condescending Language (PCL) is a subtle but harmful type of discourse, yet the task of recognizing PCL remains under-studied by the NLP community. Recognizing PCL is challenging because of its subtle nature, because available datasets are limited in size, and because this task often relies on some form of commonsense knowledge. In this paper, we study to what extent PCL detection models can be improved by pre-training them on other, more established NLP tasks. We find that performance gains are indeed possible in this way, in particular when pre-training on tasks focusing on sentiment, harmful language and commonsense morality. In contrast, for tasks focusing on political speech and social justice, no or only very small improvements were witnessed. These findings improve our understanding of the nature of PCL.},
  url   = {https://aclanthology.org/2022.lrec-1.415}
}

@InProceedings{jauhiainen-jauhiainen-lindn:2022:LREC,
  author   = {Jauhiainen, Tommi and Jauhiainen, Heidi and Lindén, Krister},
  title   = {HeLI-OTS, Off-the-shelf Language Identifier for Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3912--3922},
  abstract = {This paper introduces HeLI-OTS, an off-the-shelf text language identification tool using the HeLI language identification method. The HeLI-OTS language identifier is equipped with language models for 200 languages and licensed for academic as well as commercial use. We present the HeLI method and its use in our previous research. Then we compare the performance of the HeLI-OTS language identifier with that of fastText on two different data sets, showing that fastText favors the recall of common languages, whereas HeLI-OTS reaches both high recall and high precision for all languages. While introducing existing off-the-shelf language identification tools, we also give a picture of digital humanities-related research that uses such tools. The validity of the results of such research depends on the results given by the language identifier used, and especially for research focusing on the less common languages, the tendency to favor widely used languages might be very detrimental, which Heli-OTS is now able to remedy.},
  url   = {https://aclanthology.org/2022.lrec-1.416}
}

@InProceedings{severini-EtAl:2022:LREC,
  author   = {Severini, Silvia and ImaniGooghari, Ayyoob and Dufter, Philipp and Schütze, Hinrich},
  title   = {Towards a Broad Coverage Named Entity Resource: A Data-Efficient Approach for Many Diverse Languages},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3923--3933},
  abstract = {Parallel corpora are ideal for extracting a multilingual named entity (MNE) resource, i.e., a dataset of names translated into multiple languages. Prior work on extracting MNE datasets from parallel corpora required resources such as large monolingual corpora or word aligners that are unavailable or perform poorly for underresourced languages. We present CLC-BN, a new method for creating an MNE resource, and apply it to the Parallel Bible Corpus, a corpus of more than 1000 languages. CLC-BN learns a neural transliteration model from parallel-corpus statistics, without requiring any other bilingual resources, word aligners, or seed data. Experimental results show that CLC-BN clearly outperforms prior work. We release an MNE resource for 1340 languages and demonstrate its effectiveness in two downstream tasks: knowledge graph augmentation and bilingual lexicon induction.},
  url   = {https://aclanthology.org/2022.lrec-1.417}
}

@InProceedings{khan-EtAl:2022:LREC,
  author   = {Khan, Fahad and Minaya Gómez, Francisco J. and Cruz González, Rafael and Diakoff, Harry and Diaz Vera, Javier E. and McCrae, John P. and O'Loughlin, Ciara and Short, William Michael and Stolk, Sander},
  title   = {Towards the Construction of a WordNet for Old English},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3934--3941},
  abstract = {In this paper we will discuss our preliminary work towards the construction of a WordNet for Old English, taking our inspiration from other similar WN construction projects for ancient languages such as Ancient Greek, Latin and Sanskrit. The Old English WordNet (OldEWN) will build upon this innovative work in a number of different ways which we articulate in the article, most importantly by treateating figurative meaning as a 'first-class citizen' in the structuring of the semantic system. From a more practical perspective we will describe our plan to utilize a pre-existing lexicographic resource and the naisc system to automatically compile a provisional version of the WordNet which will then be checked and enriched by Old English experts.},
  url   = {https://aclanthology.org/2022.lrec-1.418}
}

@InProceedings{bick:2022:LREC,

  author   = {Bick, Eckhard},
  title   = {A Framenet and Frame Annotator for German Social Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3942--3949},
  abstract = {This paper presents PFN-DE, a new, parsing- and annotation-oriented framenet for German, with almost 15,000 frames, covering 11,300 verb lemmas. The resource was developed in the context of a Danish/German social-media study on hate speech and has a strong focus on coverage, robustness and cross-language comparability. A simple annotation scheme for argument roles meshes directly with the output of a syntactic parser, facilitating frame disambiguation through slot-filler conditions based on valency, syntactic function and semantic noun class. We discuss design principles for the framenet and the frame tagger using it, and present statistics for frame and role distribution at both the lexicon (type) and corpus (token) levels. In an evaluation run on Twitter data, the parser-based frame annotator achieved an overall F-score for frame senses of 93.6\%.},
  url   = {https://aclanthology.org/2022.lrec-1.419}
}

@InProceedings{koloski-EtAl:2022:LREC,
  author   = {Koloski, Boshko and Pollak, Senja and Škrlj, Blaž and Martinc, Matej},
  title   = {Out of Thin Air: Is Zero-Shot Cross-Lingual Keyword Detection Better Than Unsupervised?},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {400--409},
  abstract = {Keyword extraction is the task of retrieving words that are essential to the content of a given document. Researchers proposed various approaches to tackle this problem. At the top-most level, approaches are divided into ones that require training - supervised and ones that do not - unsupervised. In this study, we are interested in settings, where for a language under investigation, no training data is available. More specifically, we explore whether pretrained multilingual language models can be employed for zero-shot cross-lingual keyword extraction on low-resource languages with limited or no available labeled training data and whether they outperform state-of-the-art unsupervised keyword extractors. The comparison is conducted on six news article datasets covering two high-resource languages, English and Russian, and four low-resource languages, Croatian, Estonian, Latvian, and Slovenian. We find that the pretrained models fine-tuned on a

multilingual corpus covering languages that do not appear in the test set (i.e. in a zero-shot setting), consistently outscore unsupervised models in all six languages.},
  url   = {https://aclanthology.org/2022.lrec-1.42}
}

@InProceedings{bombieri-EtAl:2022:LREC,
  author   = {Bombieri, Marco and Rospocher, Marco and Ponzetto, Simone Paolo and Fiorini, Paolo},
  title   = {The Robotic Surgery Procedural Framebank},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3950--3959},
  abstract = {Robot-Assisted minimally invasive robotic surgery is the gold standard for the surgical treatment of many pathological conditions, and several manuals and academic papers describe how to perform these interventions. These high-quality, often peer-reviewed texts are the main study resource for medical personnel and consequently contain essential procedural domain-specific knowledge. The procedural knowledge therein described could be extracted, e.g., on the basis of semantic parsing models, and used to develop clinical decision support systems or even automation methods for some procedure's steps. However, natural language understanding algorithms such as, for instance, semantic role labelers have lower efficacy and coverage issues when applied to domain others than those they are typically trained on (i.e., newswire text). To overcome this problem, starting from PropBank frames, we propose a new linguistic resource specific to the robotic-surgery domain, named Robotic Surgery Procedural Framebank (RSPF). We extract from robotic-surgical texts verbs and nouns that describe surgical actions and extend PropBank frames by adding any of new lemmas, frames or role sets required to cover missing lemmas, specific frames describing the surgical significance, or new semantic roles used in procedural surgical language. Our resource is publicly available and can be used to annotate corpora in the surgical domain to train and evaluate Semantic Role Labeling (SRL) systems in a challenging fine-grained domain setting.},
  url   = {https://aclanthology.org/2022.lrec-1.420}
}

@InProceedings{weber-colunga:2022:LREC,
  author   = {Weber, Jennifer and Colunga, Eliana},
  title   = {Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {3960--3968},
  abstract = {Understanding child language development requires accurately representing children’s lexicons. However, much of the past work modeling children’s vocabulary development has utilized adult-based measures. The present investigation asks whether using corpora that captures the language input of young children more accurately represents children’s vocabulary knowledge. We present a newly-created toddler corpus that incorporates transcripts of child-directed conversations, the text of picture books written for preschoolers, and dialog from G-rated movies to approximate the language input a North American preschooler might hear. We evaluate the utility of the new corpus for modeling children’s vocabulary development by building and analyzing different semantic network models and comparing them to norms based on vocabulary norms for toddlers in this age range. More specifically, the relations between words in our semantic networks were derived from skip-gram neural networks (Word2Vec) trained on our toddler corpus or on Google news. Results revealed that the models built from the toddler corpus were more accurate at predicting toddler vocabulary growth than the adult-based corpus. These results speak to the importance of selecting a corpus that matches the population of interest.},
  url   = {https://aclanthology.org/2022.lrec-1.421}
}

@InProceedings{juniarta-EtAl:2022:LREC,
  author   = {Juniarta, Nyoman and Bonami, Olivier and Hathout, Nabil and Namer, Fiammetta and Toussaint, Yannick},
  title   = {Organizing and Improving a Database of French Word Formation Using Formal Concept Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3969--3976},
  abstract = {We apply Formal Concept Analysis (FCA) to organize and to improve the quality of Démonette2, a French derivational database, through a detection of both missing and spurious derivations in the database. We represent each derivational family as a graph. Given that the subgraph relation exists among derivational families, FCA can group families and represent them in a partially ordered set (poset). This poset is also useful for improving the database. A family is regarded as a possible anomaly (meaning that it may have missing and/or spurious derivations) if its derivational graph is almost, but not completely identical to a large number of other families.},
  url   = {https://aclanthology.org/2022.lrec-1.422}
}

@InProceedings{declerck:2022:LREC,
  author   = {Declerck, Thierry},
  title   = {Towards a new Ontology for Sign Languages},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3977--3983},
  abstract = {We present the current status of a new ontology for representing constitutive elements of Sign Languages (SL). This development emerged from investigations on how to represent multimodal lexical data in the OntoLex-Lemon framework, with the goal to publish such data in the Linguistic Linked Open Data (LLOD) cloud. While studying the literature and various sites dealing with sign languages, we saw the need to harmonise all the data categories (or features) defined and used in those sources, and to organise them in an ontology to which lexical descriptions in OntoLex-Lemon could be linked. We make the code of the first version of this ontology available, so that it can be further developed collaboratively by both the Linked Data and the SL communities},
  url   = {https://aclanthology.org/2022.lrec-1.423}
}

@InProceedings{hayashi:2022:LREC,
  author   = {Hayashi, Yoshihiko},
  title   = {Towards the Detection of a Semantic Gap in the Chain of Commonsense Knowledge Triples},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3984--3993},
  abstract = {A commonsense knowledge resource organizes common sense that is not necessarily correct all the time, but most people are expected to know or believe. Such knowledge resources have recently been actively built and utilized in artificial intelligence, particularly natural language processing. In this paper, we discuss an important but not significantly discussed the issue of semantic gaps potentially existing in a commonsense knowledge graph and propose a machine learning-based approach to detect a semantic gap that may inhibit the proper chaining of knowledge triples. In order to establish this line of research, we created a pilot dataset from ConceptNet, in which chains consisting of two adjacent triples are sampled, and the validity of each chain is human-annotated. We also devised a few baseline methods for detecting the semantic gaps and compared them in small-scale experiments. Although the experimental results suggest that the detection of semantic gaps may not be a trivial task, we achieved several insights to further push this research direction, including the potential efficacy of sense embeddings and contextualized word representations enabled by a pre-trained language model.},
  url   = {https://aclanthology.org/2022.lrec-1.424}
}

@InProceedings{brassard-EtAl:2022:LREC,
  author   = {Brassard, Ana and Heinzerling, Benjamin and
Kavumba, Pride and Inui, Kentaro},
  title   = {COPA-SSE: Semi-structured Explanations for
Commonsense Reasoning},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {3994--4000},
  abstract = {We present Semi-Structured Explanations for COPA
(COPA-SSE), a new crowdsourced dataset of 9,747 semi-structured,
English common sense explanations for Choice of Plausible
Alternatives (COPA) questions. The explanations are formatted as a
set of triple-like common sense statements with ConceptNet relations
but freely written concepts. This semi-structured format strikes a
balance between the high quality but low coverage of structured data
and the lower quality but high coverage of free-form crowdsourcing.
Each explanation also includes a set of human-given quality ratings.
With their familiar format, the explanations are geared towards
commonsense reasoners operating on knowledge graphs and serve as a
starting point for ongoing work on improving such systems. The
dataset is available at https://github.com/a-brassard/copa-sse.},
  url   = {https://aclanthology.org/2022.lrec-1.425}
}

@InProceedings{khn-mitrovi-granitzer:2022:LREC,
  author   = {Kühn, Ramona and Mitrović, Jelena and Granitzer,
Michael},
  title   = {GRhOOT: Ontology of Rhetorical Figures in German},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4001--4010},
  abstract = {GRhOOT, the German RhetOrical OnTology, is a domain
ontology of 110 rhetorical figures in the German language. The
overall goal of building an ontology of rhetorical figures in German
is not only the formal representation of different rhetorical
figures, but also allowing for their easier detection, thus
improving sentiment analysis, argument mining, detection of hate
speech and fake news, machine translation, and many other tasks in
which recognition of non-literal language plays an important role.
The challenge of building such ontologies lies in classifying the
figures and assigning adequate characteristics to group them, while
considering their distinctive features. The ontology of rhetorical
figures in the Serbian language was used as a basis for our work.
Besides transferring and extending the concepts of the Serbian
ontology, we ensured completeness and consistency by using
description logic and SPARQL queries. Furthermore, we show a

decision tree to identify figures and suggest a usage scenario on how the ontology can be utilized to collect and annotate data.},
  url   = {https://aclanthology.org/2022.lrec-1.426}
}

@InProceedings{chiarcos-fth-ionov:2022:LREC1,
  author   = {Chiarcos, Christian and Fäth, Christian and Ionov, Maxim},
  title   = {Querying a Dozen Corpora and a Thousand Years with Fintan},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4011--4021},
  abstract = {Large-scale diachronic corpus studies covering longer time periods are difficult if more than one corpus are to be consulted and, as a result, different formats and annotation schemas need to be processed and queried in a uniform, comparable and replicable manner. We describes the application of the Flexible Integrated Transformation and Annotation eNgineering (Fintan) platform for studying word order in German using syntactically annotated corpora that represent its entire written history. Focusing on nominal dative and accusative arguments, this study hints at two major phases in the development of scrambling in modern German. Against more recent assumptions, it supports the traditional view that word order flexibility decreased over time, but it also indicates that this was a relatively sharp transition in Early New High German. The successful case study demonstrates the potential of Fintan and the underlying LLOD technology for historical linguistics, linguistic typology and corpus linguistics. The technological contribution of this paper is to demonstrate the applicability of Fintan for querying across heterogeneously annotated corpora, as previously, it had only been applied for transformation tasks. With its focus on quantitative analysis, Fintan is a natural complement for existing multi-layer technologies that focus on query and exploration.},
  url   = {https://aclanthology.org/2022.lrec-1.427}
}

@InProceedings{mambrini-EtAl:2022:LREC,
  author   = {Mambrini, Francesco and Passarotti, Marco and Moretti, Giovanni and Pellegrini, Matteo},
  title   = {The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4022--4029},

  abstract = {Although the Universal Dependencies initiative today allows for cross-linguistically consistent annotation of morphology and syntax in treebanks for several languages, syntactically annotated corpora are not yet interoperable with many lexical resources that describe properties of the words that occur therein. In order to cope with such limitation, we propose to adopt the principles of the Linguistic Linked Open Data community, to describe and publish dependency treebanks as LLOD. In particular, this paper illustrates the approach pursued in the LiLa Knowledge Base, which enables interoperability between corpora and lexical resources for Latin, to publish as Linguistic Linked Open Data the annotation layers of two versions of a Medieval Latin treebank (the Index Thomisticus Treebank).},
  url   = {https://aclanthology.org/2022.lrec-1.428}
}

@InProceedings{menini-EtAl:2022:LREC,
  author   = {Menini, Stefano and Paccosi, Teresa and Tekiroğlu, Serra Sinem and Tonelli, Sara},
  title   = {Building a Multilingual Taxonomy of Olfactory Terms with Timestamps},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4030--4039},
  abstract = {Olfactory references play a crucial role in our memory and, more generally, in our experiences, since researchers have shown that smell is the sense that is most directly connected with emotions. Nevertheless, only few works in NLP have tried to capture this sensory dimension from a computational perspective. One of the main challenges is the lack of a systematic and consistent taxonomy of olfactory information, where concepts are organised also in a multi-lingual perspective. WordNet represents a valuable starting point in this direction, which can be semi-automatically extended taking advantage of Google n-grams and of existing language models. In this work we describe the process that has led to the semi-automatic development of a taxonomy for olfactory information in four languages (English, French, German and Italian), detailing the different steps and the intermediate evaluations. Along with being multi-lingual, the taxonomy also encloses temporal marks for olfactory terms thus making it a valuable resource for historical content analysis. The resource has been released and is freely available.},
  url   = {https://aclanthology.org/2022.lrec-1.429}
}

@InProceedings{lamproudis-henriksson-dalianis:2022:LREC,
  author   = {Lamproudis, Anastasios and Henriksson, Aron and Dalianis, Hercules},
  title   = {Evaluating Pretraining Strategies for Clinical BERT Models},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {410--416},
  abstract = {Research suggests that using generic language models in specialized domains may be sub-optimal due to significant domain differences. As a result, various strategies for developing domain-specific language models have been proposed, including techniques for adapting an existing generic language model to the target domain, e.g. through various forms of vocabulary modifications and continued domain-adaptive pretraining with in-domain data. Here, an empirical investigation is carried out in which various strategies for adapting a generic language model to the clinical domain are compared to pretraining a pure clinical language model. Three clinical language models for Swedish, pretrained for up to ten epochs, are fine-tuned and evaluated on several downstream tasks in the clinical domain. A comparison of the language models' downstream performance over the training epochs is conducted. The results show that the domain-specific language models outperform a general-domain language model; however, there is little difference in performance of the various clinical language models. However, compared to pretraining a pure clinical language model with only in-domain data, leveraging and adapting an existing general-domain language model requires fewer epochs of pretraining with in-domain data.},
  url   = {https://aclanthology.org/2022.lrec-1.43}
}

@InProceedings{chizhikova-EtAl:2022:LREC,
  author   = {Chizhikova, Anastasia and Murzakhmetov, Sanzhar and Serikov, Oleg and Shavrina, Tatiana and Burtsev, Mikhail},
  title   = {Attention Understands Semantic Relations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4040--4050},
  abstract = {Today, natural language processing heavily relies on pre-trained large language models. Even though such models are criticized for the poor interpretability, they still yield state-of-the-art solutions for a wide set of very different tasks. While lots of probing studies have been conducted to measure the models' awareness of grammatical knowledge, semantic probing is less popular. In this work, we introduce the probing pipeline to study the representedness of semantic relations in transformer language models. We show that in this task, attention scores are nearly as expressive as the layers’ output activations, despite their lesser ability to represent surface cues. This supports the hypothesis that attention mechanisms are focusing not only on the syntactic relational information but also on the semantic one.},

  url   = {https://aclanthology.org/2022.lrec-1.430}
}

@InProceedings{ichikawa-higashinaka:2022:LREC,
  author   = {Ichikawa, Takuma and Higashinaka, Ryuichiro},
  title   = {Analysis of Dialogue in Human-Human Collaboration in Minecraft},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4051--4059},
  abstract = {Recently, many studies have focused on developing dialogue systems that enable collaborative work; however, they rarely focus on creative tasks. Collaboration for creative work, in which humans and systems collaborate to create new value, will be essential for future dialogue systems. In this study, we collected 500 dialogues of human-human collaboration in Minecraft as a basis for developing a dialogue system that enables creative collaborative work. We conceived the Collaborative Garden Task, where two workers interact and collaborate in Minecraft to create a garden, and we collected dialogue, action logs, and subjective evaluations. We also collected third-person evaluations of the gardens and analyzed the relationship between dialogue and collaborative work that received high scores on the subjective and third-person evaluations in order to identify dialogic factors for high-quality collaborative work. We found that two essential aspects in creative collaborative work are performing more processes to ask for and agree on suggestions between workers and agreeing on a particular image of the final product in the early phase of work and then discussing changes and details.},
  url   = {https://aclanthology.org/2022.lrec-1.431}
}

@InProceedings{yamashita-higashinaka:2022:LREC,
  author   = {Yamashita, Sanae and Higashinaka, Ryuichiro},
  title   = {Data Collection for Empirically Determining the Necessary Information for Smooth Handover in Dialogue},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4060--4068},
  abstract = {Despite recent advances, dialogue systems still struggle to achieve fully autonomous transactions. Therefore, when a system encounters a problem, human operators need to take over the dialogue to complete the transaction. However, it is unclear what information should be presented to the operator when this handover takes place. In this study, we conducted a data collection experiment in which one of two operators talked to a user and

switched with the other operator periodically while exchanging notes when the handovers took place. By examining these notes, it is possible to identify the information necessary for handing over the dialogue. We collected 60 dialogues in which two operators switched periodically while performing chat, consultation, and sales tasks in dialogue. We found that adjacency pairs are a useful representation for recording conversation history. In addition, we found that key-value-pair representation is also useful when there are underlying tasks, such as consultation and sales.},
  url   = {https://aclanthology.org/2022.lrec-1.432}
}

@InProceedings{gtze-EtAl:2022:LREC,
  author   = {Götze, Jana and Paetzel-Prüsmann, Maike and Liermann, Wencke and Diekmann, Tim and Schlangen, David},
  title   = {The slurk Interaction Server Framework: Better Data for Better Dialog Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4069--4078},
  abstract = {This paper presents the slurk software, a lightweight interaction server for setting up dialog data collections and running experiments. slurk enables a multitude of settings including text-based, speech and video interaction between two or more humans or humans and bots, and a multimodal display area for presenting shared or private interactive context. The software is implemented in Python with an HTML and JavaScript frontend that can easily be adapted to individual needs. It also provides a setup for pairing participants on common crowdworking platforms such as Amazon Mechanical Turk and some example bot scripts for common interaction scenarios.},
  url   = {https://aclanthology.org/2022.lrec-1.433}
}

@InProceedings{kalashnikova-EtAl:2022:LREC,
  author   = {Kalashnikova, Natalia and Pajak, Serge and Le Guel, Fabrice and Vasilescu, Ioana and Serrano, Gemma and Devillers, Laurence},
  title   = {Corpus Design for Studying Linguistic Nudges in Human-Computer Spoken Interactions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4079--4087},
  abstract = {In this paper, we present the methodology of corpus design that will be used to study the comparison of influence between linguistic nudges with positive or negative influences and

three conversational agents: robot, smart speaker, and human. We recruited forty-nine participants to form six groups. The conversational agents first asked the participants about their willingness to adopt five ecological habits and invest time and money in ecological problems. The participants were then asked the same questions but preceded by one linguistic nudge with positive or negative influence. The comparison of standard deviation and mean metrics of differences between these two notes (before the nudge and after) showed that participants were mainly affected by nudges with positive influence, even though several nudges with negative influence decreased the average note. In addition, participants from all groups were willing to spend more money than time on ecological problems. In general, our experiment's early results suggest that a machine agent can influence participants to the same degree as a human agent. A better understanding of the power of influence of different conversational machines and the potential of influence of nudges of different polarities will lead to the development of ethical norms of human-computer interactions.},
  url   = {https://aclanthology.org/2022.lrec-1.434}
}

@InProceedings{furuya-EtAl:2022:LREC,
  author   = {Furuya, Yuki and Saito, Koki and Ogura, Kosuke and Mitsuda, Koh and Higashinaka, Ryuichiro and Takashio, Kazunori},
  title   = {Dialogue Corpus Construction Considering Modality and Social Relationships in Building Common Ground},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4088--4095},
  abstract = {Building common ground with users is essential for dialogue agent systems and robots to interact naturally with people. While a few previous studies have investigated the process of building common ground in human-human dialogue, most of them have been conducted on the basis of text chat. In this study, we constructed a dialogue corpus to investigate the process of building common ground with a particular focus on the modality of dialogue and the social relationship between the participants in the process of building common ground, which are important but have not been investigated in the previous work. The results of our analysis suggest that adding the modality or developing the relationship between workers speeds up the building of common ground. Specifically, regarding the modality, the presence of video rather than only audio may unconsciously facilitate work, and as for the relationship, it is easier to convey information about emotions and turn-taking among friends than in first meetings. These findings and the corpus should prove useful for developing a system to support remote communication.},
  url   = {https://aclanthology.org/2022.lrec-1.435}
}

@InProceedings{feng-EtAl:2022:LREC,
  author   = {Feng, Shutong and Lubis, Nurul and Geishauser, Christian and Lin, Hsien-chin and Heck, Michael and van Niekerk, Carel and Gasic, Milica},
  title   = {EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4096--4113},
  abstract = {The ability to recognise emotions lends a conversational artificial intelligence a human touch. While emotions in chit-chat dialogues have received substantial attention, emotions in task-oriented dialogues remain largely unaddressed. This is despite emotions and dialogue success having equally important roles in a natural system. Existing emotion-annotated task-oriented corpora are limited in size, label richness, and public availability, creating a bottleneck for downstream tasks. To lay a foundation for studies on emotions in task-oriented dialogues, we introduce EmoWOZ, a large-scale manually emotion-annotated corpus of task-oriented dialogues. EmoWOZ is based on MultiWOZ, a multi-domain task-oriented dialogue dataset. It contains more than 11K dialogues with more than 83K emotion annotations of user utterances. In addition to Wizard-of-Oz dialogues from MultiWOZ, we collect human-machine dialogues within the same set of domains to sufficiently cover the space of various emotions that can happen during the lifetime of a data-driven dialogue system. To the best of our knowledge, this is the first large-scale open-source corpus of its kind. We propose a novel emotion labelling scheme, which is tailored to task-oriented dialogues. We report a set of experimental results to show the usability of this corpus for emotion recognition and state tracking in task-oriented dialogues.},
  url   = {https://aclanthology.org/2022.lrec-1.436}
}

@InProceedings{okur-sahay-nachman:2022:LREC,
  author   = {Okur, Eda and Sahay, Saurav and Nachman, Lama},
  title   = {Data Augmentation with Paraphrase Generation and Entity Extraction for Multimodal Dialogue System},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4114--4125},
  abstract = {Contextually aware intelligent agents are often required to understand the users and their surroundings in real-time. Our goal is to build Artificial Intelligence (AI) systems that can assist children in their learning process. Within such complex

frameworks, Spoken Dialogue Systems (SDS) are crucial building blocks to handle efficient task-oriented communication with children in game-based learning settings. We are working towards a multimodal dialogue system for younger kids learning basic math concepts. Our focus is on improving the Natural Language Understanding (NLU) module of the task-oriented SDS pipeline with limited datasets. This work explores the potential benefits of data augmentation with paraphrase generation for the NLU models trained on small task-specific datasets. We also investigate the effects of extracting entities for conceivably further data expansion. We have shown that paraphrasing with model-in-the-loop (MITL) strategies using small seed data is a promising approach yielding improved performance results for the Intent Recognition task.},
  url   = {https://aclanthology.org/2022.lrec-1.437}
}

@InProceedings{aicher-minker-ultes:2022:LREC,
  author   = {Aicher, Annalena and Minker, Wolfgang and Ultes, Stefan},
  title   = {Towards Modelling Self-imposed Filter Bubbles in Argumentative Dialogue Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4126--4134},
  abstract = {To build a well-founded opinion it is natural for humans to gather and exchange new arguments. Especially when being confronted with an overwhelming amount of information, people tend to focus on only the part of the available information that fits into their current beliefs or convenient opinions. To overcome this "self-imposed filter bubble'' (SFB) in the information seeking process, it is crucial to identify influential indicators for the former. Within this paper we propose and investigate indicators for the the user's SFB, mainly their Reflective User Engagement (RUE), their Personal Relevance (PR) ranking of content-related subtopics as well as their False (FK) and True Knowledge (TK) on the topic. Therefore, we analysed the answers of 202 participants of an online conducted user study, who interacted with our argumentative dialogue system BEA ("Building Engaging Argumentation''). Moreover, also the influence of different input/output modalities (speech/speech and drop-down menu/text) on the interaction with regard to the suggested indicators was investigated.},
  url   = {https://aclanthology.org/2022.lrec-1.438}
}

@InProceedings{aich-parde:2022:LREC,
  author   = {Aich, Ankit and Parde, Natalie},
  title   = {Telling a Lie: Analyzing the Language of Information and Misinformation during Global Health Events},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4135--4141},
  abstract = {The COVID-19 pandemic and other global health events are unfortunately excellent environments for the creation and spread of misinformation, and the language associated with health misinformation may be typified by unique patterns and linguistic markers. Allowing health misinformation to spread unchecked can have devastating ripple effects; however, detecting and stopping its spread requires careful analysis of these linguistic characteristics at scale. We analyze prior investigations focusing on health misinformation, associated datasets, and detection of misinformation during health crises. We also introduce a novel dataset designed for analyzing such phenomena, comprised of 2.8 million news articles and social media posts spanning the early 1900s to the present. Our annotation guidelines result in strong agreement between independent annotators. We describe our methods for collecting this data and follow this with a thorough analysis of the themes and linguistic features that appear in information versus misinformation. Finally, we demonstrate a proof-of-concept misinformation detection task to establish dataset validity, achieving a strong performance benchmark (accuracy = 75\%; F1 = 0.7).},
  url   = {https://aclanthology.org/2022.lrec-1.439}
}

@InProceedings{yeshpanov-khassanov-varol:2022:LREC,
  author   = {Yeshpanov, Rustem and Khassanov, Yerbolat and Varol, Huseyin Atakan},
  title   = {KazNERD: Kazakh Named Entity Recognition Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {417--426},
  abstract = {We present the development of a dataset for Kazakh named entity recognition. The dataset was built as there is a clear need for publicly available annotated corpora in Kazakh, as well as annotation guidelines containing straightforward—but rigorous—rules and examples. The dataset annotation, based on the IOB2 scheme, was carried out on television news text by two native Kazakh speakers under the supervision of the first author. The resulting dataset contains 112,702 sentences and 136,333 annotations for 25 entity classes. State-of-the-art machine learning models to automatise Kazakh named entity recognition were also built, with the best-performing model achieving an exact match F1-score of 97.22\% on the test set. The annotated dataset, guidelines, and codes used to train the models are freely available for download under the CC BY 4.0 licence from https://github.com/IS2AI/KazNERD.},
  url   = {https://aclanthology.org/2022.lrec-1.44}
}

@InProceedings{muti-fernicola-barrncedeo:2022:LREC,
  author   = {Muti, Arianna and Fernicola, Francesco and Barrón-Cedeño, Alberto},
  title   = {Misogyny and Aggressiveness Tend to Come Together and Together We Address Them},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4142--4148},
  abstract = {We target the complementary binary tasks of identifying whether a tweet is misogynous and, if that is the case, whether it is also aggressive. We compare two ways to address these problems: one multi-class model that discriminates between all the classes at once: not misogynous, non aggressive-misogynous and aggressive-misogynous; as well as a cascaded approach where the binary classification is carried out separately (misogynous vs non-misogynous and aggressive vs non-aggressive) and then joined together. For the latter, two training and three testing scenarios are considered. Our models are built on top of AlBERTo and are evaluated on the framework of Evalita's 2020 shared task on automatic misogyny and aggressiveness identification in Italian tweets. Our cascaded models ---including the strong naïve baseline--- outperform significantly the top submissions to Evalita, reaching state-of-the-art performance without relying on any external information.},
  url   = {https://aclanthology.org/2022.lrec-1.440}
}

@InProceedings{kumar-EtAl:2022:LREC1,
  author   = {Kumar, Ritesh and Ratan, Shyam and Singh, Siddharth and Nandi, Enakshi and Devi, Laishram Niranjana and Bhagat, Akash and Dawer, Yogesh and lahiri, bornini and Bansal, Akanksha and Ojha, Atul Kr.},
  title   = {The ComMA Dataset V0.2: Annotating Aggression and Bias in Multilingual Social Media Discourse},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4149--4161},
  abstract = {In this paper, we discuss the development of a multilingual dataset annotated with a hierarchical, fine-grained tagset marking different types of aggression and the "context" in which they occur. The context, here, is defined by the conversational thread in which a specific comment occurs and also the "type" of discursive role that the comment is performing with respect to the previous comment. The initial dataset, being discussed here consists of a total 59,152 annotated comments in four

languages - Meitei, Bangla, Hindi, and Indian English - collected from various social media platforms such as YouTube, Facebook, Twitter and Telegram. As is usual on social media websites, a large number of these comments are multilingual, mostly code-mixed with English. The paper gives a detailed description of the tagset being used for annotation and also the process of developing a multi-label, fine-grained tagset that has been used for marking comments with aggression and bias of various kinds including sexism (called gender bias in the tagset), religious intolerance (called communal bias in the tagset), class/caste bias and ethnic/racial bias. We also define and discuss the tags that have been used for marking the different discursive role being performed through the comments, such as attack, defend, etc. Finally, we present a basic statistical analysis of the dataset. The dataset is being incrementally made publicly available on the project website\footnote{\url{https://sites.google.com/view/comma-ctrans}}.},
  url   = {https://aclanthology.org/2022.lrec-1.441}
}

@InProceedings{vishnubhotla-mohammad:2022:LREC,
  author   = {Vishnubhotla, Krishnapriya and Mohammad, Saif M.},
  title   = {TUSC: Emotion Word Usage in Tweets from US and Canada},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4162--4176},
  abstract = {Over the last decade, Twitter has emerged as one of the most influential forums for social, political, and health discourse. In this paper, we introduce a massive dataset of more than 45 million geo-located tweets posted between 2015 and 2021 from US and Canada (TUSC), especially curated for natural language analysis. We also introduce Tweet Emotion Dynamics (TED) - metrics to capture patterns of emotions associated with tweets over time. We use TED and TUSC to explore the use of emotion-associated words across US and Canada; across 2019 (pre-pandemic), 2020 (the year the pandemic hit), and 2021 (the second year of the pandemic); and across individual tweeters. We show that Canadian tweets tend to have higher valence, lower arousal, and higher dominance than the US tweets. Further, we show that the COVID-19 pandemic had a marked impact on the emotional signature of tweets posted in 2020, when compared to the adjoining years. Finally, we determine metrics of TED for 170,000 tweeters to benchmark characteristics of TED metrics at an aggregate level. TUSC and the metrics for TED will enable a wide variety of research on studying how we use language to express ourselves, persuade, communicate, and influence, with particularly promising applications in public health, affective science, social science, and psychology.},
  url   = {https://aclanthology.org/2022.lrec-1.442}
}

@InProceedings{beyhan-EtAl:2022:LREC,
  author   = {Beyhan, Fatih and Çarık, Buse and Arın, İnanç and Terzioğlu, Ayşecan and Yanikoglu, Berrin and Yeniterzi, Reyyan},
  title   = {A Turkish Hate Speech Dataset and Detection System},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4177--4185},
  abstract = {Social media posts containing hate speech are reproduced and redistributed at an accelerated pace, reaching greater audiences at a higher speed. We present a machine learning system for automatic detection of hate speech in Turkish, along with a hate speech dataset consisting of tweets collected in two separate domains. We first adopted a definition for hate speech that is in line with our goals and amenable to easy annotation; then designed the annotation schema for annotating the collected tweets. The Istanbul Convention dataset consists of tweets posted following the withdrawal of Turkey from the Istanbul Convention. The Refugees dataset was created by collecting tweets about immigrants by filtering based on commonly used keywords related to immigrants. Finally, we have developed a hate speech detection system using the transformer architecture (BERTurk), to be used as a baseline for the collected dataset. The binary classification accuracy is 77\% when the system is evaluated using 5-fold cross-validation on the Istanbul Convention dataset and 71\% for the Refugee dataset. We also tested a regression model with 0.66 and 0.83 RMSE on a scale of [0-4], for the Istanbul Convention and Refugees datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.443}
}

@InProceedings{bucur-cosma-dinu:2022:LREC,
  author   = {Bucur, Ana-Maria and Cosma, Adrian and Dinu, Liviu P.},
  title   = {Life is not Always Depressing: Exploring the Happy Moments of People Diagnosed with Depression},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4186--4192},
  abstract = {In this work, we explore the relationship between depression and manifestations of happiness in social media. While the majority of works surrounding depression focus on symptoms, psychological research shows that there is a strong link between seeking happiness and being diagnosed with depression. We make use of Positive-Unlabeled learning paradigm to automatically extract happy moments from social media posts of both controls and users diagnosed with depression, and qualitatively analyze them with

linguistic tools such as LIWC and keyness information. We show that the life of depressed individuals is not always bleak, with positive events related to friends and family being more noteworthy to their lives compared to the more mundane happy events reported by control users.},
  url   = {https://aclanthology.org/2022.lrec-1.444}
}

@InProceedings{benamar-EtAl:2022:LREC,
  author   = {Benamar, Alexandra and Grouin, Cyril and Bothua, Meryl and Vilnat, Anne},
  title   = {Evaluating Tokenizers Impact on OOVs Representation with Transformers Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4193--4204},
  abstract = {Transformer models have achieved significant improvements in multiple downstream tasks in recent years. One of the main contributions of Transformers is their ability to create new representations for out-of-vocabulary (OOV) words. In this paper, we have evaluated three categories of OOVs: (A) new domain-specific terms (e.g., "eucaryote'" in microbiology), (B) misspelled words containing typos, and (C) cross-domain homographs (e.g., "arm'' has different meanings in a clinical trial and anatomy). We use three French domain-specific datasets on the legal, medical, and energetical domains to robustly analyze these categories. Our experiments have led to exciting findings that showed: (1) It is easier to improve the representation of new words (A and B) than it is for words that already exist in the vocabulary of the Transformer models (C), (2) To ameliorate the representation of OOVs, the most effective method relies on adding external morpho-syntactic context rather than improving the semantic understanding of the words directly (fine-tuning) and (3) We cannot foresee the impact of minor misspellings in words because similar misspellings have different impacts on their representation. We believe that tackling the challenges of processing OOVs regarding their specificities will significantly help the domain adaptation aspect of BERT.},
  url   = {https://aclanthology.org/2022.lrec-1.445}
}

@InProceedings{morza-manna-monti:2022:LREC,
  author   = {Morza, Giuseppina and Manna, Raffaele and Monti, Johanna},
  title   = {Assessing the Quality of an Italian Crowdsourced Idiom Corpus:the Dodiom Experiment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {4205--4211},
  abstract = {This paper describes how idiom-related language resources, collected through a crowdsourcing experiment carried out by means of Dodiom, a Game-with-a-purpose, have been analysed by language experts. The paper focuses on the criteria adopted for the data annotation and evaluation process. The main scope of this project is, indeed, the evaluation of the quality of the linguistic data obtained through a crowdsourcing project, namely to assess if the data provided and evaluated by the players who joined the game are actually considered of good quality by the language experts. Finally, results of the annotation and evaluation processes as well as future work are presented.},
  url   = {https://aclanthology.org/2022.lrec-1.446}
}

@InProceedings{alekseev-EtAl:2022:LREC,
  author   = {Alekseev, Anton and Miftahutdinov, Zulfat and Tutubalina, Elena and Shelmanov, Artem and Ivanov, Vladimir and Kokh, Vladimir and Nesterov, Alexander and Avetisian, Manvel and Chertok, Andrei and Nikolenko, Sergey},
  title   = {Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4212--4220},
  abstract = {Medical data annotation requires highly qualified expertise. Despite the efforts devoted to medical entity linking in different languages, available data is very sparse in terms of both data volume and languages. In this work, we establish benchmarks for cross-lingual medical entity linking using clinical reports, clinical guidelines, and medical research papers. We present a test set filtering procedure designed to analyze the ``hard cases'' of entity linking approaching zero-shot cross-lingual transfer learning, evaluate state-of-the-art models, and draw several interesting conclusions based on our evaluation results.},
  url   = {https://aclanthology.org/2022.lrec-1.447}
}

@InProceedings{sharma-EtAl:2022:LREC,
  author   = {Sharma, Shreyas and Darwish, Kareem and Pavanelli, Lucas and Castro Ferreira, Thiago and Al-Badrashiny, Mohamed and Yuksel, Kamer Ali and Sawaf, Hassan},
  title   = {MTLens: Machine Translation Output Debugging},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {4221--4226},
  abstract = {The performance of Machine Translation (MT) systems varies significantly with inputs of diverging features such as topics, genres, and surface properties. Though there are many MT evaluation metrics that generally correlate with human judgments, they are not directly useful in identifying specific shortcomings of MT systems. In this demo, we present a benchmarking interface that enables improved evaluation of specific MT systems in isolation or multiple MT systems collectively by quantitatively evaluating their performance on many tasks across multiple domains and evaluation metrics. Further, it facilitates effective debugging and error analysis of MT output via the use of dynamic filters that help users hone in on problem sentences with specific properties, such as genre, topic, sentence length, etc. The interface can be extended to include additional filters such as lexical, morphological, and syntactic features. Aside from helping debug MT output, it can also help in identifying problems in reference translations and evaluation metrics.},
  url   = {https://aclanthology.org/2022.lrec-1.448}
}

@InProceedings{fririksdttir-EtAl:2022:LREC,
  author   = {Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór and Sigurdsson, Einar},
  title   = {IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4227--4234},
  abstract = {Word embedding models have become commonplace in a wide range of NLP applications. In order to train and use the best possible models, accurate evaluation is needed. For extrinsic evaluation of word embedding models, analogy evaluation sets have been shown to be a good quality estimator. We introduce an Icelandic adaptation of a large analogy dataset, BATS, evaluate it on three different word embedding models and show that our evaluation set is apt at measuring the capabilities of such models.},
  url   = {https://aclanthology.org/2022.lrec-1.449}
}

@InProceedings{mersinias-valvis:2022:LREC,
  author   = {Mersinias, Michail and Valvis, Panagiotis},
  title   = {Mitigating Dataset Artifacts in Natural Language Inference Through Automatic Contextual Data Augmentation and Learning Optimization},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

publisher = {European Language Resources Association},
pages = {427--435},
abstract = {In recent years, natural language inference has been an emerging research area. In this paper, we present a novel data augmentation technique and combine it with a unique learning procedure for that task. Our so-called automatic contextual data augmentation (acda) method manages to be fully automatic, non-trivially contextual, and computationally efficient at the same time. When compared to established data augmentation methods, it is substantially more computationally efficient and requires no manual annotation by a human expert as they usually do. In order to increase its efficiency, we combine acda with two learning optimization techniques: contrastive learning and a hybrid loss function. The former maximizes the benefit of the supervisory signal generated by acda, while the latter incentivises the model to learn the nuances of the decision boundary. Our combined approach is shown experimentally to provide an effective way for mitigating spurious data correlations within a dataset, called dataset artifacts, and as a result improves performance. Specifically, our experiments verify that acda-boosted pre-trained language models that employ our learning optimization techniques, consistently outperform the respective fine-tuned baseline pre-trained language models across both benchmark datasets and adversarial examples.},
url = {https://aclanthology.org/2022.lrec-1.45}
}

@InProceedings{akhbardeh-EtAl:2022:LREC,
author = {Akhbardeh, Farhad and Zampieri, Marcos and Alm, Cecilia Ovesdotter and Desell, Travis},
title = {Transfer Learning Methods for Domain Adaptation in Technical Logbook Datasets},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {4235--4244},
abstract = {Event identification in technical logbooks poses challenges given the limited logbook data available in specific technical domains, the large set of possible classes, and logbook entries typically being in short form and non-standard technical language. Technical logbook data typically has both a domain, the field it comes from (e.g., automotive), and an application, what it is used for (e.g., maintenance). In order to better handle the problem of data scarcity, using a variety of technical logbook datasets, this paper investigates the benefits of using transfer learning from sources within the same domain (but different applications), from within the same application (but different domains) and from all available data. Results show that performing transfer learning within a domain provides statistically significant improvements, and in all cases but one the best performance. Interestingly, transfer learning from within the application or across the global dataset degrades results in all cases but one,

which benefited from adding as much data as possible. A further analysis of the dataset similarities shows that the datasets with higher similarity scores performed better in transfer learning tasks, suggesting that this can be utilized to determine the effectiveness of adding a dataset in a transfer learning task for technical logbooks.},
  url   = {https://aclanthology.org/2022.lrec-1.450}
}

@InProceedings{vakili-EtAl:2022:LREC,
  author   = {Vakili, Thomas and Lamproudis, Anastasios and Henriksson, Aron and Dalianis, Hercules},
  title   = {Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4245--4252},
  abstract = {Automatic de-identification is a cost-effective and straightforward way of removing large amounts of personally identifiable information from large and sensitive corpora. However, these systems also introduce errors into datasets due to their imperfect precision. These corruptions of the data may negatively impact the utility of the de-identified dataset. This paper de-identifies a very large clinical corpus in Swedish either by removing entire sentences containing sensitive data or by replacing sensitive words with realistic surrogates. These two datasets are used to perform domain adaptation of a general Swedish BERT model. The impact of the de-identification techniques is assessed by training and evaluating the models using six clinical downstream tasks. The results are then compared to a similar BERT model domain-adapted using an unaltered version of the clinical corpus. The results show that using an automatically de-identified corpus for domain adaptation does not negatively impact downstream performance. We argue that automatic de-identification is an efficient way of reducing the privacy risks of domain-adapted models and that the models created in this paper should be safe to distribute to other academic researchers.},
  url   = {https://aclanthology.org/2022.lrec-1.451}
}

@InProceedings{csandy-lukcs:2022:LREC,
  author   = {Csanády, Bálint and Lukács, András},
  title   = {Dilated Convolutional Neural Networks for Lightweight Diacritics Restoration},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {4253--4259},
  abstract = {Diacritics restoration has become a ubiquitous task in the Latin-alphabet-based English-dominated Internet language environment. In this paper, we describe a small footprint 1D dilated convolution-based approach which operates on a character-level. We find that neural networks based on 1D dilated convolutions are competitive alternatives to solutions based on recurrent neural networks or linguistic modeling for the task of diacritics restoration. Our approach surpasses the performance of similarly sized models and is also competitive with larger models. A special feature of our solution is that it even runs locally in a web browser. We also provide a working example of this browser-based implementation. Our model is evaluated on different corpora, with emphasis on the Hungarian language. We performed comparative measurements about the generalization power of the model in relation to three Hungarian corpora. We also analyzed the errors to understand the limitation of corpus-based self-supervised training.},
  url   = {https://aclanthology.org/2022.lrec-1.452}
}

@InProceedings{claveau-chaffin-kijak:2022:LREC,
  author   = {Claveau, Vincent and Chaffin, Antoine and Kijak, Ewa},
  title   = {Generating Artificial Texts as Substitution or Complement of Training Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4260--4269},
  abstract = {The quality of artificially generated texts has considerably improved with the advent of transformers. The question of using these models to generate learning data for supervised learning tasks naturally arises, especially when the original language resource cannot be distributed, or when it is small. In this article, this question is explored under 3 aspects: (i) are artificial data an efficient complement? (ii) can they replace the original data when those are not available or cannot be distributed for confidentiality reasons? (iii) can they improve the explainability of classifiers? Different experiments are carried out on classification tasks - namely sentiment analysis on product reviews and Fake News detection - using artificially generated data by fine-tuned GPT-2 models. The results show that such artificial data can be used in a certain extend but require pre-processing to significantly improve performance. We also show that bag-of-words approaches benefit the most from such data augmentation.},
  url   = {https://aclanthology.org/2022.lrec-1.453}
}

@InProceedings{coeckelbergs:2022:LREC,
  author   = {Coeckelbergs, Mathias},

  title   = {From Pattern to Interpretation. Using Colibri Core to Detect Translation Patterns in the Peshitta.},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4270--4274},
  abstract = {This article presents the first results of the CLARIAH-funded project ‘Patterns in Translation: Using Colibri Core for the Syriac Bible’ (PaTraCoSy). This project seeks to use Colibri Core to detect translation patterns in the Peshitta, the Syriac translation of the Hebrew Bible. We first describe how we constructed word and phrase alignment between these two texts. This step is necessary to succesfully implement the functionalities of Colibri Core. After this, we further describe our first investigations with the software. We describe how we use the built-in pattern modeller to detect n-gram and skipgram patterns in both Hebrew and Syriac texts. Colibri Core does not allow the creation of a bilingual model, which is why we compare the separate models. After a presentation of a few general insights on the overall translation behaviour of the Peshitta, we delve deeper into the concrete patterns we can detect by the n-gram/skipgram analysis. We provide multiple examples from the book of Genesis, a book which has been treated broadly in scholarly research into the Syriac translation, but which also appears to have interesting features based on our Colibri Core research.},
  url   = {https://aclanthology.org/2022.lrec-1.454}
}

@InProceedings{launay-EtAl:2022:LREC,
  author   = {Launay, Julien and Tommasone, E.L. and Pannier, Baptiste and Boniface, François and Chatelain, Amélie and Cappelli, Alessandro and Poli, Iacopo and Seddah, Djamé},
  title   = {PAGnol: An Extra-Large French Generative Model},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4275--4284},
  abstract = {Access to large pre-trained models of varied architectures, in many different languages, is central to the democratization of NLP. We introduce PAGnol, a collection of French GPT models. Using scaling laws, we efficiently train PAGnol-XL (1.5B parameters) with the same computational budget as CamemBERT, a model 13 times smaller. PAGnol-XL is the largest model trained from scratch for the French language. We plan to train increasingly large and performing versions of PAGnol, exploring the capabilities of French extreme-scale models. For this first release, we focus on the pre-training and scaling calculations underlining PAGnol. We fit a scaling law for compute for the French language, and compare it with

its English counterpart. We find the pre-training dataset significantly conditions the quality of the outputs, with common datasets such as OSCAR leading to low-quality offensive text. We evaluate our models on discriminative and generative tasks in French, comparing to other state-of-the-art French and multilingual models, and reaching the state of the art in the abstract summarization task. Our research was conducted on the public GENCI Jean Zay supercomputer, and our models up to the Large are made publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.455}
}

@InProceedings{felice-EtAl:2022:LREC,
  author   = {Felice, Mariano and Taslimipoor, Shiva and Andersen, Øistein E. and Buttery, Paula},
  title   = {CEPOC: The Cambridge Exams Publishing Open Cloze dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4285--4290},
  abstract = {Open cloze tests are a standard type of exercise where examinees must complete a text by filling in the gaps without any given options to choose from. This paper presents the Cambridge Exams Publishing Open Cloze (CEPOC) dataset, a collection of open cloze tests from world-renowned English language proficiency examinations. The tests in CEPOC have been expertly designed and validated using standard principles in language research and assessment. They are prepared for language learners at different proficiency levels and hence classified into different CEFR levels (A2, B1, B2, C1, C2). This resource can be a valuable testbed for various NLP tasks. We perform a complete set of experiments on three tasks: gap filling, gap prediction, and CEFR text classification. We implement transformer-based systems based on pre-trained language models to model each task and use our dataset as a test set, providing promising benchmark results.},
  url   = {https://aclanthology.org/2022.lrec-1.456}
}

@InProceedings{caete-EtAl:2022:LREC,
  author   = {Cañete, José and Donoso, Sebastian and Bravo-Marquez, Felipe and Carvallo, Andrés and Araujo, Vladimir},
  title   = {ALBETO and DistilBETO: Lightweight Spanish Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4291--4298},

  abstract = {In recent years there have been considerable advances in pre-trained language models, where non-English language versions have also been made available. Due to their increasing use, many lightweight versions of these models (with reduced parameters) have also been released to speed up training and inference times. However, versions of these lighter models (e.g., ALBERT, DistilBERT) for languages other than English are still scarce. In this paper we present ALBETO and DistilBETO, which are versions of ALBERT and DistilBERT pre-trained exclusively on Spanish corpora. We train several versions of ALBETO ranging from 5M to 223M parameters and one of DistilBETO with 67M parameters. We evaluate our models in the GLUES benchmark that includes various natural language understanding tasks in Spanish. The results show that our lightweight models achieve competitive results to those of BETO (Spanish-BERT) despite having fewer parameters. More specifically, our larger ALBETO model outperforms all other models on the MLDoc, PAWS-X, XNLI, MLQA, SQAC and XQuAD datasets. However, BETO remains unbeaten for POS and NER. As a further contribution, all models are publicly available to the community for future research.},
  url   = {https://aclanthology.org/2022.lrec-1.457}
}

@InProceedings{wu-yarowsky:2022:LREC,
  author   = {Wu, Winston and Yarowsky, David},
  title   = {On the Robustness of Cognate Generation Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4299--4305},
  abstract = {We evaluate two popular neural cognate generation models' robustness to several types of human-plausible noise (deletion, duplication, swapping, and keyboard errors, as well as a new type of error, phonological errors). We find that duplication and phonological substitution is least harmful, while the other types of errors are harmful. We present an in-depth analysis of the models' results with respect to each error type to explain how and why these models perform as they do.},
  url   = {https://aclanthology.org/2022.lrec-1.458}
}

@InProceedings{hiebel-EtAl:2022:LREC,
  author   = {Hiebel, Nicolas and Ferret, Olivier and Fort, Karën and Névéol, Aurélie},
  title   = {CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {4306--4315},
  abstract = {Modern Natural Language Processing relies on the availability of annotated corpora for training and evaluating models. Such resources are scarce, especially for specialized domains in languages other than English. In particular, there are very few resources for semantic similarity in the clinical domain in French. This can be useful for many biomedical natural language processing applications, including text generation. We introduce a definition of similarity that is guided by clinical facts and apply it to the development of a new French corpus of 1,000 sentence pairs manually annotated according to similarity scores. This new sentence similarity corpus is made freely available to the community. We further evaluate the corpus through experiments of automatic similarity measurement. We show that a model of sentence embeddings can capture similarity with state-of-the-art performance on the DEFT STS shared task evaluation data set (Spearman=0.8343). We also show that the \corpus corpus is complementary to DEFT STS.},
  url   = {https://aclanthology.org/2022.lrec-1.459}
}

@InProceedings{zhang-jensen-plank:2022:LREC,
  author   = {Zhang, Mike and Jensen, Kristian Nørgaard and Plank, Barbara},
  title   = {Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {436--447},
  abstract = {Skill Classification (SC) is the task of classifying job competences from job postings. This work is the first in SC applied to Danish job vacancy data. We release the first Danish job posting dataset: *Kompetencer* (_en_: competences), annotated for nested spans of competences. To improve upon coarse-grained annotations, we make use of The European Skills, Competences, Qualifications and Occupations (ESCO; le Vrang et al., (2014)) taxonomy API to obtain fine-grained labels via distant supervision. We study two setups: The zero-shot and few-shot classification setting. We fine-tune English-based models and RemBERT (Chung et al., 2020) and compare them to in-language Danish models. Our results show RemBERT significantly outperforms all other models in both the zero-shot and the few-shot setting.},
  url   = {https://aclanthology.org/2022.lrec-1.46}
}

@InProceedings{xu-markert:2022:LREC,
  author   = {Xu, Shanshan and Markert, Katja},
  title   = {The Chinese Causative-Passive Homonymy Disambiguation: an adversarial Dataset for NLI and a Probing Task},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4316--4323},
  abstract = {The disambiguation of causative-passive homonymy (CPH) is potentially tricky for machines, as the causative and the passive are not distinguished by the sentences' syntactic structure. By transforming CPH disambiguation to a challenging natural language inference (NLI) task, we present the first Chinese Adversarial NLI challenge set (CANLI). We show that the pretrained transformer model RoBERTa, fine-tuned on an existing large-scale Chinese NLI benchmark dataset, performs poorly on CANLI. We also employ Word Sense Disambiguation as a probing task to investigate to what extent the CPH feature is captured in the model's internal representation. We find that the model's performance on CANLI does not correspond to its internal representation of CPH, which is the crucial linguistic ability central to the CANLI dataset. CANLI is available on Hugging Face Datasets (Lhoest et al., 2021) at https://huggingface.co/datasets/sxu/CANLI},
  url   = {https://aclanthology.org/2022.lrec-1.460}
}

@InProceedings{vahtola-EtAl:2022:LREC,
  author   = {Vahtola, Teemu and Sjöblom, Eetu and Tiedemann, Jörg and Creutz, Mathias},
  title   = {Modeling Noise in Paraphrase Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4324--4332},
  abstract = {Noisy labels in training data present a challenging issue in classification tasks, misleading a model towards incorrect decisions during training. In this paper, we propose the use of a linear noise model to augment pre-trained language models to account for label noise in fine-tuning. We test our approach in a paraphrase detection task with various levels of noise and five different languages. Our experiments demonstrate the effectiveness of the additional noise model in making the training procedures more robust and stable. Furthermore, we show that this model can be applied without further knowledge about annotation confidence and reliability of individual training examples and we analyse our results in light of data selection and sampling strategies.},
  url   = {https://aclanthology.org/2022.lrec-1.461}
}

@InProceedings{laurenti-EtAl:2022:LREC,
  author   = {laurenti, Enzo and Bourgon, Nils and Benamara, Farah and Mari, Alda and MORICEAU, Véronique and Courgeon, Camille},
  title   = {Give me your Intentions, I'll Predict our Actions: A

Two-level Classification of Speech Acts for Crisis Management in Social Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4333--4343},
  abstract = {Discovered by (Austin,1962) and extensively promoted by (Searle, 1975), speech acts (SA) have been the object of extensive discussion in the philosophical and the linguistic literature, as well as in computational linguistics where the detection of SA have shown to be an important step in many down stream NLP applications. In this paper, we attempt to measure for the first time the role of SA on urgency detection in tweets, focusing on natural disasters. Indeed, SA are particularly relevant to identify intentions, desires, plans and preferences towards action, providing therefore actionable information that will help to set priorities for the human teams and decide appropriate rescue actions. To this end, we come up here with four main contributions: (1) A two-layer annotation scheme of SA both at the tweet and subtweet levels, (2) A new French dataset of 6,669 tweets annotated for both urgency and SA, (3) An in-depth analysis of the annotation campaign, highlighting the correlation between SA and urgency categories, and (4) A set of deep learning experiments to detect SA in a crisis corpus. Our results show that SA are correlated with urgency which is a first important step towards SA-aware NLP-based crisis management on social media.},
  url   = {https://aclanthology.org/2022.lrec-1.462}
}

@InProceedings{abadji-EtAl:2022:LREC,
  author   = {Abadji, Julien and Ortiz Suarez, Pedro and Romary, Laurent and Sagot, Benoît},
  title   = {Towards a Cleaner Document-Oriented Multilingual Crawled Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4344--4355},
  abstract = {The need for large corpora raw corpora has dramatically increased in recent years with the introduction of transfer learning and semi-supervised learning methods to Natural Language Processing. And while there have been some recent attempts to manually curate the amount of data necessary to train large language models, the main way to obtain this data is still through automatic web crawling. In this paper we take the existing multilingual web corpus OSCAR and its pipeline Ungoliant that extracts and classifies data from Common Crawl at the line level, and propose a set of improvements and automatic annotations in order

to produce a new document-oriented version of OSCAR that could prove more suitable to pre-train large generative language models as well as hopefully other applications in Natural Language Processing and Digital Humanities.},
  url   = {https://aclanthology.org/2022.lrec-1.463}
}

@InProceedings{snbjarnarson-EtAl:2022:LREC,
  author   = {Snæbjarnarson, Vésteinn and Símonarson, Haukur Barri and Ragnarsson, Pétur Orri and Ingólfsdóttir, Svanhvít Lilja and Jónsson, Haukur and Thorsteinsson, Vilhjalmur and Einarsson, Hafsteinn},
  title   = {A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4356--4366},
  abstract = {We train several language models for Icelandic, including IceBERT, that achieve state-of-the-art performance in a variety of downstream tasks, including part-of-speech tagging, named entity recognition, grammatical error detection and constituency parsing. To train the models we introduce a new corpus of Icelandic text, the Icelandic Common Crawl Corpus (IC3), a collection of high quality texts found online by targeting the Icelandic top-level-domain .is. Several other public data sources are also collected for a total of 16GB of Icelandic text. To enhance the evaluation of model performance and to raise the bar in baselines for Icelandic, we manually translate and adapt the WinoGrande commonsense reasoning dataset. Through these efforts we demonstrate that a properly cleaned crawled corpus is sufficient to achieve state-of-the-art results in NLP applications for low to medium resource languages, by comparison with models trained on a curated corpus. We further show that initializing models using existing multilingual models can lead to state-of-the-art results for some downstream tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.464}
}

@InProceedings{turan-EtAl:2022:LREC,
  author   = {Turan, Tugtekin and Klakow, Dietrich and Vincent, Emmanuel and Jouvet, Denis},
  title   = {Adapting Language Models When Training on Privacy-Transformed Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4367--4373},
  abstract = {In recent years, voice-controlled personal assistants

have revolutionized the interaction with smart devices and mobile applications. The collected data are then used by system providers to train language models (LMs). Each spoken message reveals personal information, hence removing private information from the input sentences is necessary. Our data sanitization process relies on recognizing and replacing named entities by other words from the same class. However, this may harm LM training because privacy-transformed data is unlikely to match the test distribution. This paper aims to fill the gap by focusing on the adaptation of LMs initially trained on privacy-transformed sentences using a small amount of original untransformed data. To do so, we combine class-based LMs, which provide an effective approach to overcome data sparsity in the context of n-gram LMs, and neural LMs, which handle longer contexts and can yield better predictions. Our experiments show that training an LM on privacy-transformed data result in a relative 11\% word error rate (WER) increase compared to training on the original untransformed data, and adapting that model on a limited amount of original untransformed data leads to a relative 8\% WER improvement over the model trained solely on privacy-transformed data.},
  url = {https://aclanthology.org/2022.lrec-1.465}
}

@InProceedings{chrabrowa-EtAl:2022:LREC,
  author = {Chrabrowa, Aleksandra and Dragan, Łukasz and Grzegorczyk, Karol and Kajtoch, Dariusz and Koszowski, Mikołaj and Mroczkowski, Robert and Rybak, Piotr},
  title = {Evaluation of Transfer Learning for Polish with a Text-to-Text Model},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {4374--4394},
  abstract = {We introduce a new benchmark for assessing the quality of text-to-text models for Polish. The benchmark consists of diverse tasks and datasets: KLEJ benchmark adapted for text-to-text, en-pl translation, summarization, and question answering. In particular, since summarization and question answering lack benchmark datasets for the Polish language, we describe in detail their construction and make them publicly available. Additionally, we present plT5 - a general-purpose text-to-text model for Polish that can be fine-tuned on various Natural Language Processing (NLP) tasks with a single training objective. Unsupervised denoising pre-training is performed efficiently by initializing the model weights with a multi-lingual T5 (mT5) counterpart. We evaluate the performance of plT5, mT5, Polish BART (plBART), and Polish GPT-2 (papuGaPT2). The plT5 scores top on all of these tasks except summarization, where plBART is best. In general (except summarization), the larger the model, the better the results. The encoder-decoder architectures prove to be better than the decoder-only equivalent.},

  url   = {https://aclanthology.org/2022.lrec-1.466}
}

@InProceedings{strbel-EtAl:2022:LREC,
  author   = {Ströbel, Phillip Benjamin and Volk, Martin and Clematide, Simon and Schwitter, Raphael and Hodel, Tobias and Schoch, David},
  title   = {Evaluation of HTR models without Ground Truth Material},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4395--4404},
  abstract = {The evaluation of Handwritten Text Recognition (HTR) models during their development is straightforward: because HTR is a supervised problem, the usual data split into training, validation, and test data sets allows the evaluation of models in terms of accuracy or error rates. However, the evaluation process becomes tricky as soon as we switch from development to application. A compilation of a new (and forcibly smaller) ground truth (GT) from a sample of the data that we want to apply the model on and the subsequent evaluation of models thereon only provides hints about the quality of the recognised text, as do confidence scores (if available) the models return. Moreover, if we have several models at hand, we face a model selection problem since we want to obtain the best possible result during the application phase. This calls for GT-free metrics to select the best model, which is why we (re-)introduce and compare different metrics, from simple, lexicon-based to more elaborate ones using standard language models and masked language models (MLM). We show that MLM-based evaluation can compete with lexicon-based methods, with the advantage that large and multilingual transformers are readily available, thus making compiling lexical resources for other metrics superfluous.},
  url   = {https://aclanthology.org/2022.lrec-1.467}
}

@InProceedings{korybski-EtAl:2022:LREC,
  author   = {Korybski, Tomasz and Davitti, Elena and Orasan, Constantin and Braun, Sabine},
  title   = {A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4405--4413},
  abstract = {In this paper, we present a semi-automated workflow for live interlingual speech-to-text communication which seeks to

reduce the shortcomings of existing ASR systems: a human respeaker works with a speaker-dependent speech recognition software (e.g., Dragon Naturally Speaking) to deliver punctuated same-language output of superior quality than obtained using out-of-the-box automatic speech recognition of the original speech. This is fed into a machine translation engine (the EU's eTranslation) to produce live-caption ready text. We benchmark the quality of the output against the output of best-in-class (human) simultaneous interpreters working with the same source speeches from plenary sessions of the European Parliament. To evaluate the accuracy and facilitate the comparison between the two types of output, we use a tailored annotation approach based on the NTR model (Romero-Fresco and Pöchhacker, 2017). We find that the semi-automated workflow combining intralingual respeaking and machine translation is capable of generating outputs that are similar in terms of accuracy and completeness to the outputs produced in the benchmarking workflow, although the small scale of our experiment requires caution in interpreting this result.},
  url   = {https://aclanthology.org/2022.lrec-1.468}
}

@InProceedings{prouteau-EtAl:2022:LREC,
  author   = {Prouteau, Thibault and Dugué, Nicolas and
Camelin, Nathalie and Meignier, Sylvain},
  title   = {Are Embedding Spaces Interpretable? Results of an
Intrusion Detection Evaluation on a Large French Corpus},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4414--4419},
  abstract = {Word embedding methods allow to represent words as
vectors in a space that is structured using word co-occurrences so
that words with close meanings are close in this space. These
vectors are then provided as input to automatic systems to solve
natural language processing problems. Because interpretability is a
necessary condition to trusting such systems, interpretability of
embedding spaces, the first link in the chain is an important issue.
In this paper, we thus evaluate the interpretability of vectors
extracted with two approaches: SPINE a k-sparse auto-encoder, and
SINr, a graph-based method. This evaluation is based on a Word
Intrusion Task with human annotators. It is operated using a large
French corpus, and is thus, as far as we know, the first large-scale
experiment regarding word embedding interpretability on this
language. Furthermore, contrary to the approaches adopted in the
literature where the evaluation is done on a small sample of
frequent words, we consider a more realistic use-case where most of
the vocabulary is kept for the evaluation. This allows to show how
difficult this task is, even though SPINE and SINr show some
promising results. In particular, SINr results are obtained with a
very low amount of computation compared to SPINE, while being
similarly interpretable.},

  url   = {https://aclanthology.org/2022.lrec-1.469}
}

@InProceedings{bakker-EtAl:2022:LREC,
  author   = {Bakker, Roos and van Drie, Romy A.N. and de Boer, Maaike and van Doesburg, Robert and van Engers, Tom},
  title   = {Semantic Role Labelling for Dutch Law Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {448--457},
  abstract = {Legal texts are often difficult to interpret, and people who interpret them need to make choices about the interpretation. To improve transparency, the interpretation of a legal text can be made explicit by formalising it. However, creating formalised representations of legal texts manually is quite labour-intensive. In this paper, we describe a method to extract structured representations in the Flint language (van Doesburg and van Engers, 2019) from natural language. Automated extraction of knowledge representation not only makes the interpretation and modelling efforts more efficient, it also contributes to reducing inter-coder dependencies. The Flint language offers a formal model that enables the interpretation of legal text by describing the norms in these texts as acts, facts and duties. To extract the components of a Flint representation, we use a rule-based method and a transformer-based method. In the transformer-based method we fine-tune the last layer with annotated legal texts. The results show that the transformed-based method (80\% accuracy) outperforms the rule-based method (42\% accuracy) on the Dutch Aliens Act. This indicates that the transformer-based method is a promising approach of automatically extracting Flint frames.},
  url   = {https://aclanthology.org/2022.lrec-1.47}
}

@InProceedings{kalamkar-EtAl:2022:LREC,
  author   = {Kalamkar, Prathamesh and Tiwari, Aman and Agarwal, Astha and Karn, Saurabh and Gupta, Smita and Raghavan, Vivek and Modi, Ashutosh},
  title   = {Corpus for Automatic Structuring of Legal Documents},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4420--4429},
  abstract = {In populous countries, pending legal cases have been growing exponentially. There is a need for developing techniques for processing and organizing legal documents. In this paper, we introduce a new corpus for structuring legal documents. In particular, we introduce a corpus of legal judgment documents in

English that are segmented into topical and coherent parts. Each of these parts is annotated with a label coming from a list of pre-defined Rhetorical Roles. We develop baseline models for automatically predicting rhetorical roles in a legal document based on the annotated corpus. Further, we show the application of rhetorical roles to improve performance on the tasks of summarization and legal judgment prediction. We release the corpus and baseline model code along with the paper.},
  url   = {https://aclanthology.org/2022.lrec-1.470}
}

@InProceedings{bonial-EtAl:2022:LREC,
  author   = {Bonial, Claire and Blodgett, Austin and Hudson, Taylor and Lukin, Stephanie M. and Micher, Jeffrey and Summers-Stay, Douglas and Sutor, Peter and Voss, Clare},
  title   = {The Search for Agreement on Logical Fallacy Annotation of an Infodemic},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4430--4438},
  abstract = {We evaluate an annotation schema for labeling logical fallacy types, originally developed for a crowd-sourcing annotation paradigm, now using an annotation paradigm of two trained linguist annotators. We apply the schema to a variety of different genres of text relating to the COVID-19 pandemic. Our linguist (as opposed to crowd-sourced) annotation of logical fallacies allows us to evaluate whether the annotation schema category labels are sufficiently clear and non-overlapping for both manual and, later, system assignment. We report inter-annotator agreement results over two annotation phases as well as a preliminary assessment of the corpus for training and testing a machine learning algorithm (Pattern-Exploiting Training) for fallacy detection and recognition. The agreement results and system performance underscore the challenging nature of this annotation task and suggest that the annotation schema and paradigm must be iteratively evaluated and refined in order to arrive at a set of annotation labels that can be reproduced by human annotators and, in turn, provide reliable training data for automatic detection and recognition systems.},
  url   = {https://aclanthology.org/2022.lrec-1.471}
}

@InProceedings{whrl-klinger:2022:LREC,
  author   = {Wührl, Amelie and Klinger, Roman},
  title   = {Recovering Patient Journeys: A Corpus of Biomedical Entities and Relations on Twitter (BEAR)},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {4439--4450},
  abstract = {Text mining and information extraction for the medical domain has focused on scientific text generated by researchers. However, their access to individual patient experiences or patient-doctor interactions is limited. On social media, doctors, patients and their relatives also discuss medical information. Individual information provided by laypeople complements the knowledge available in scientific text. It reflects the patient's journey making the value of this type of data twofold: It offers direct access to people's perspectives, and it might cover information that is not available elsewhere, including self-treatment or self-diagnose. Named entity recognition and relation extraction are methods to structure information that is available in unstructured text. However, existing medical social media corpora focused on a comparably small set of entities and relations. In contrast, we provide rich annotation layers to model patients' experiences in detail. The corpus consists of medical tweets annotated with a fine-grained set of medical entities and relations between them, namely 14 entity (incl. environmental factors, diagnostics, biochemical processes, patients' quality-of-life descriptions, pathogens, medical conditions, and treatments) and 20 relation classes (incl. prevents, influences, interactions, causes). The dataset consists of 2,100 tweets with approx. 6,000 entities and 2,200 relations.},
  url   = {https://aclanthology.org/2022.lrec-1.472}
}

@InProceedings{virgo-cheng-kurohashi:2022:LREC,
  author   = {Virgo, Felix and Cheng, Fei and Kurohashi, Sadao},
  title   = {Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4451--4457},
  abstract = {Understanding event duration is essential for understanding natural language. However, the amount of training data for tasks like duration question answering, i.e., McTACO, is very limited, suggesting a need for external duration information to improve this task. The duration information can be obtained from existing temporal information extraction tasks, such as UDS-T and TimeBank, where more duration data is available. A straightforward two-stage fine-tuning approach might be less likely to succeed given the discrepancy between the target duration question answering task and the intermediary duration classification task. This paper resolves this discrepancy by automatically recasting an existing event duration classification task from UDS-T to a question answering task similar to the target McTACO. We investigate the transferability of duration information by comparing whether the

original UDS-T duration classification or the recast UDS-T duration question answering can be transferred to the target task. Our proposed model achieves a 13\% Exact Match score improvement over the baseline on the McTACO duration question answering task, showing that the two-stage fine-tuning approach succeeds when the discrepancy between the target and intermediary tasks are resolved.},
  url   = {https://aclanthology.org/2022.lrec-1.473}
}

@InProceedings{loukachevitch-EtAl:2022:LREC,
  author   = {Loukachevitch, Natalia and Braslavski, Pavel and Ivanov, Vladimir and Batura, Tatiana and Manandhar, Suresh and Shelmanov, Artem and Tutubalina, Elena},
  title   = {Entity Linking over Nested Named Entities for Russian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4458--4466},
  abstract = {In this paper, we describe entity linking annotation over nested named entities in the recently released Russian NEREL dataset for information extraction. The NEREL collection is currently the largest Russian dataset annotated with entities and relations. It includes 933 news texts with annotation of 29 entity types and 49 relation types. The paper describes the main design principles behind NEREL's entity linking annotation, provides its statistics, and reports evaluation results for several entity linking baselines. To date, 38,152 entity mentions in 933 documents are linked to Wikidata. The NEREL dataset is publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.474}
}

@InProceedings{murthy-EtAl:2022:LREC,
  author   = {Murthy, Rudra and Bhattacharjee, Pallab and Sharnagat, Rahul and Khatri, Jyotsana and Kanojia, Diptesh and Bhattacharyya, Pushpak},
  title   = {HiNER: A large Hindi Named Entity Recognition Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4467--4476},
  abstract = {Named Entity Recognition (NER) is a foundational NLP task that aims to provide class labels like Person, Location, Organisation, Time, and Number to words in free text. Named Entities can also be multi-word expressions where the additional I-O-B annotation information helps label them during the NER annotation

process. While English and European languages have considerable annotated data for the NER task, Indian languages lack on that front- both in terms of quantity and following annotation standards. This paper releases a significantly sized standard-abiding Hindi NER dataset containing 109,146 sentences and 2,220,856 tokens, annotated with 11 tags. We discuss the dataset statistics in all their essential detail and provide an in-depth analysis of the NER tag-set used with our data. The statistics of tag-set in our dataset shows a healthy per-tag distribution especially for prominent classes like Person, Location and Organisation. Since the proof of resource-effectiveness is in building models with the resource and testing the model on benchmark data and against the leader-board entries in shared tasks, we do the same with the aforesaid data. We use different language models to perform the sequence labelling task for NER and show the efficacy of our data by performing a comparative evaluation with models trained on another dataset available for the Hindi NER task. Our dataset helps achieve a weighted F1 score of 88.78 with all the tags and 92.22 when we collapse the tag-set, as discussed in the paper. To the best of our knowledge, no available dataset meets the standards of volume (amount) and variability (diversity), as far as Hindi NER is concerned. We fill this gap through this work, which we hope will significantly help NLP for Hindi. We release this dataset with our code and models for further research at https://github.com/cfiltnlp/HiNER},
  url   = {https://aclanthology.org/2022.lrec-1.475}
}

@InProceedings{papadopoulou-EtAl:2022:LREC,
  author   = {Papadopoulou, Anthi and Lison, Pierre and Øvrelid, Lilja and Pilán, Ildikó},
  title   = {Bootstrapping Text Anonymization Models with Distant Supervision},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4477--4487},
  abstract = {We propose a novel method to bootstrap text anonymization models based on distant supervision. Instead of requiring manually labeled training data, the approach relies on a knowledge graph expressing the background information assumed to be publicly available about various individuals. This knowledge graph is employed to automatically annotate text documents including personal data about a subset of those individuals. More precisely, the method determines which text spans ought to be masked in order to guarantee k-anonymity, assuming an adversary with access to both the text documents and the background information expressed in the knowledge graph. The resulting collection of labeled documents is then used as training data to fine-tune a pre-trained language model for text anonymization. We illustrate this approach using a knowledge graph extracted from Wikidata and short biographical texts from Wikipedia. Evaluation results with a RoBERTa-based model and a

manually annotated collection of 553 summaries showcase the potential of the approach, but also unveil a number of issues that may arise if the knowledge graph is noisy or incomplete. The results also illustrate that, contrary to most sequence labeling problems, the text anonymization task may admit several alternative solutions.},
  url   = {https://aclanthology.org/2022.lrec-1.476}
}

@InProceedings{snbjarnarson-einarsson:2022:LREC,
  author   = {Snæbjarnarson, Vésteinn and Einarsson, Hafsteinn},
  title   = {Natural Questions in Icelandic},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4488--4496},
  abstract = {We present the first extractive question answering (QA) dataset for Icelandic, Natural Questions in Icelandic (NQiI). Developing such datasets is important for the development and evaluation of Icelandic QA systems. It also aids in the development of QA methods that need to work for a wide range of morphologically and grammatically different languages in a multilingual setting. The dataset was created by asking contributors to come up with questions they would like to know the answer to. Later, they were tasked with finding answers to each others questions following a previously published methodology. The questions are Natural in the sense that they are real questions posed out of interest in knowing the answer. The complete dataset contains 18 thousand labeled entries of which 5,568 are directly suitable for training an extractive QA system for Icelandic. The dataset is a valuable resource for Icelandic which we demonstrate by creating and evaluating a system capable of extractive QA in Icelandic.},
  url   = {https://aclanthology.org/2022.lrec-1.477}
}

@InProceedings{silva-EtAl:2022:LREC,
  author   = {Silva, Rafael Jimenez and Gedela, Kaushik and Marr, Alex and Desmet, Bart and Rose, Carolyn and Zhou, Chunxiao},
  title   = {QA4IE: A Quality Assurance Tool for Information Extraction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4497--4503},
  abstract = {Quality assurance (QA) is an essential though underdeveloped part of the data annotation process. Although QA is supported to some extent in existing annotation tools, comprehensive

support for QA is not standardly provided. In this paper we contribute QA4IE, a comprehensive QA tool for information extraction, which can (1) detect potential problems in text annotations in a timely manner, (2) accurately assess the quality of annotations, (3) visually display and summarize annotation discrepancies among annotation team members, (4) provide a comprehensive statistics report, and (5) support viewing of annotated documents interactively. This paper offers a competitive analysis comparing QA4IE and other popular annotation tools and demonstrates its features, usage, and effectiveness through a case study. The Python code, documentation, and demonstration video are available publicly at https://github.com/CC-RMD-EpiBio/QA4IE.},
  url   = {https://aclanthology.org/2022.lrec-1.478}
}

@InProceedings{schirmer-kruschwitz-donabauer:2022:LREC,
  author   = {Schirmer, Miriam and Kruschwitz, Udo and Donabauer, Gregor},
  title   = {A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4504--4512},
  abstract = {Recent progress in natural language processing has been impressive in many different areas with transformer-based approaches setting new benchmarks for a wide range of applications. This development has also lowered the barriers for people outside the NLP community to tap into the tools and resources applied to a variety of domain-specific applications. The bottleneck however still remains the lack of annotated gold-standard collections as soon as one's research or professional interest falls outside the scope of what is readily available. One such area is genocide-related research (also including the work of experts who have a professional interest in accessing, exploring and searching large-scale document collections on the topic, such as lawyers). We present GTC (Genocide Transcript Corpus), the first annotated corpus of genocide-related court transcripts which serves three purposes: (1) to provide a first reference corpus for the community, (2) to establish benchmark performances (using state-of-the-art transformer-based approaches) for the new classification task of paragraph identification of violence-related witness statements, (3) to explore first steps towards transfer learning within the domain. We consider our contribution to be addressing in particular this year's hot topic on Language Technology for All.},
  url   = {https://aclanthology.org/2022.lrec-1.479}
}

@InProceedings{goslin-hofmann:2022:LREC,
  author   = {Goslin, Kyle and Hofmann, Markus},
  title   = {English Language Spelling Correction as an

Information Retrieval Task Using Wikipedia Search Statistics},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {458--464},
  abstract = {Spelling correction utilities have become commonplace during the writing process, however, many spelling correction utilities suffer due to the size and quality of dictionaries available to aid correction. Many terms, acronyms, and morphological variations of terms are often missing, leaving potential spelling errors unidentified and potentially uncorrected. This research describes the implementation of WikiSpell, a dynamic spelling correction tool that relies on the Wikipedia dataset search API functionality as the sole source of knowledge to aid misspelled term identification and automatic replacement. Instead of a traditional matching process to select candidate replacement terms, the replacement process is treated as a natural language information retrieval process harnessing wildcard string matching and search result statistics. The aims of this research include: 1) the implementation of a spelling correction algorithm that utilizes the wildcard operators in the Wikipedia dataset search API, 2) a review of the current spell correction tools and approaches being utilized, and 3) testing and validation of the developed algorithm against the benchmark spelling correction tool, Hunspell. The key contribution of this research is a robust, dynamic information retrieval-based spelling correction algorithm that does not require prior training. Results of this research show that the proposed spelling correction algorithm, WikiSpell, achieved comparable results to an industry-standard spelling correction algorithm, Hunspell.},
  url   = {https://aclanthology.org/2022.lrec-1.48}
}

@InProceedings{nascimento-EtAl:2022:LREC,
  author   = {Nascimento, Igor and Lima, Rinaldo and CHIFU, Adrian-Gabriel and Espinasse, Bernard and Fournier, Sébastien},
  title   = {DeepREF: A Framework for Optimized Deep Learning-based Relation Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4513--4522},
  abstract = {The Relation Extraction (RE) is an important basic Natural Language Processing (NLP) for many applications, such as search engines, recommender systems, question-answering systems and others. There are many studies in this subarea of NLP that continue to be explored, such as SemEval campaigns (2010 to 2018), or DDI Extraction (2013).For more than ten years, different RE systems using mainly statistical models have been proposed as well as the

frameworks to develop them. This paper focuses on frameworks allowing to develop such RE systems using deep learning models. Such frameworks should make it possible to reproduce experiments of various deep learning models and pre-processing techniques proposed in various publications. Currently, there are very few frameworks of this type, and we propose a new open and optimizable framework, called DeepREF, which is inspired by the OpenNRE and REflex existing frameworks. DeepREF allows the employment of various deep learning models, to optimize their use, to identify the best inputs and to get better results with each data set for RE and compare with other experiments, making ablation studies possible. The DeepREF Framework is evaluated on several reference corpora from various application domains.},
  url   = {https://aclanthology.org/2022.lrec-1.480}
}

@InProceedings{azam-rizwan-karim:2022:LREC,
  author   = {Azam, Ubaid and Rizwan, Hammad and Karim, Asim},
  title   = {Exploring Data Augmentation Strategies for Hate Speech Detection in Roman Urdu},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4523--4531},
  abstract = {In an era where social media platform users are growing rapidly, there has been a marked increase in hateful content being generated; to combat this, automatic hate speech detection systems are a necessity. For this purpose, researchers have recently focused their efforts on developing datasets, however, the vast majority of them have been generated for the English language, with only a few available for low-resource languages such as Roman Urdu. Furthermore, what few are available have small number of samples that pertain to hateful classes and these lack variations in topics and content. Thus, deep learning models trained on such datasets perform poorly when deployed in the real world. To improve performance the option of collecting and annotating more data can be very costly and time consuming. Thus, data augmentation techniques need to be explored to exploit already available datasets to improve model generalizability. In this paper, we explore different data augmentation techniques for the improvement of hate speech detection in Roman Urdu. We evaluate these augmentation techniques on two datasets. We are able to improve performance in the primary metric of comparison (F1 and Macro F1) as well as in recall, which is impertinent for human-in-the-loop AI systems.},
  url   = {https://aclanthology.org/2022.lrec-1.481}
}

@InProceedings{yakutkilic-pan:2022:LREC,
  author   = {Yakut Kilic, Isil and Pan, Shimei},
  title   = {Incorporating LIWC in Neural Networks to Improve Human Trait and Behavior Analysis in Low Resource Scenarios},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4532--4539},
  abstract = {Psycholinguistic knowledge resources have been widely used in constructing features for text-based human trait and behavior analysis. Recently, deep neural network (NN)-based text analysis methods have gained dominance due to their high prediction performance. However, NN-based methods may not perform well in low resource scenarios where the ground truth data is limited (e.g., only a few hundred labeled training instances are available). In this research, we investigate diverse methods to incorporate Linguistic Inquiry and Word Count (LIWC), a widely-used psycholinguistic lexicon, in NN models to improve human trait and behavior analysis in low resource scenarios. We evaluate the proposed methods in two tasks: predicting delay discounting and predicting drug use based on social media posts. The results demonstrate that our methods perform significantly better than baselines that use only LIWC or only NN-based feature learning methods. They also performed significantly better than published results on the same dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.482}
}

@InProceedings{mullick-EtAl:2022:LREC1,
  author   = {Mullick, Ankan and Pal, Shubhraneel and Nayak, Tapas and Lee, Seung-Cheol and Bhattacharjee, Satadeep and Goyal, Pawan},
  title   = {Using Sentence-level Classification Helps Entity Extraction from Material Science Literature},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4540--4545},
  abstract = {In the last few years, several attempts have been made on extracting information from material science research domain. Material Science research articles are a rich source of information about various entities related to material science such as names of the materials used for experiments, the computational software used along with its parameters, the method used in the experiments, etc. But the distribution of these entities is not uniform across different sections of research articles. Most of the sentences in the research articles do not contain any entity. In this work, we first use a sentence-level classifier to identify sentences containing at least one entity mention. Next, we apply the information extraction models only on the filtered sentences, to extract various entities of interest. Our experiments for named entity recognition in the material science research articles show

that this additional sentence-level classification step helps to improve the F1 score by more than 4\%.},
  url   = {https://aclanthology.org/2022.lrec-1.483}
}

@InProceedings{ark-yeniterzi:2022:LREC,
  author   = {Çarık, Buse and Yeniterzi, Reyyan},
  title   = {A Twitter Corpus for Named Entity Recognition in Turkish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4546--4551},
  abstract = {This paper introduces a new Turkish Twitter Named Entity Recognition dataset. The dataset, which consists of 5000 tweets from a year-long period, was labeled by multiple annotators with a high agreement score. The dataset is also diverse in terms of the named entity types as it contains not only person, organization, and location but also time, money, product, and tv-show categories. Our initial experiments with pretrained language models (like BertTurk) over this dataset returned F1 scores of around 80\%. We share this dataset publicly.},
  url   = {https://aclanthology.org/2022.lrec-1.484}
}

@InProceedings{luo-surdeanu:2022:LREC,
  author   = {Luo, Fan and Surdeanu, Mihai},
  title   = {A STEP towards Interpretable Multi-Hop Reasoning:Bridge Phrase Identification and Query Expansion},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4552--4560},
  abstract = {We propose an unsupervised method for the identification of bridge phrases in multi-hop question answering (QA). Our method constructs a graph of noun phrases from the question and the available context, and applies the Steiner tree algorithm to identify the minimal sub-graph that connects all question phrases. Nodes in the sub-graph that bridge loosely-connected or disjoint subsets of question phrases due to low-strength semantic relations are extracted as bridge phrases. The identified bridge phrases are then used to expand the query based on the initial question, helping in increasing the relevance of evidence that has little lexical overlap or semantic relation with the question. Through an evaluation on HotpotQA, a popular dataset for multi-hop QA, we show that our method yields: (a) improved evidence retrieval, (b) improved QA performance when using the retrieved sentences; and (c) effective and faithful explanations

when answers are provided.},
  url   = {https://aclanthology.org/2022.lrec-1.485}
}

@InProceedings{bechet-EtAl:2022:LREC,
  author   = {Bechet, Frederic and Antoine, Elie and Auguste, Jérémy and Damnati, Géraldine},
  title   = {Question Generation and Answering for exploring Digital Humanities collections},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4561--4568},
  abstract = {This paper introduces the question answering paradigm as a way to explore digitized archive collections for Social Science studies. In particular, we are interested in evaluating largely studied question generation and question answering approaches on a new type of documents, as a step forward beyond traditional benchmark evaluations. Question generation can be used as a way to provide enhanced training material for Machine Reading Question Answering algorithms but also has its own purpose in this paradigm, where relevant questions can be used as a way to create explainable links between documents. To this end, generating large amounts of question is not the only motivation, but we need to include qualitative and semantic control to the generation process. We propose a new approach for question generation, relying on a BART Transformer based generative model, for which input data are enriched by semantic constraints. Question generation and answering are evaluated on several French corpora, and the whole approach is validated on a new corpus of digitized archive collection of a French Social Science journal.},
  url   = {https://aclanthology.org/2022.lrec-1.486}
}

@InProceedings{ide-EtAl:2022:LREC,
  author   = {Ide, Nancy and Suderman, Keith and Tu, Jingxuan and Verhagen, Marc and Peters, Shanan and Ross, Ian and Lawson, John and Borg, Andrew and Pustejovsky, James},
  title   = {Evaluating Retrieval for Multi-domain Scientific Publications},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4569--4576},
  abstract = {This paper provides an overview of the xDD/LAPPS Grid framework and provides results of evaluating the AskMe retrievalengine using the BEIR benchmark datasets. Our primary goal is to determine a solid baseline of performance to guide

furtherdevelopment of our retrieval capabilities. Beyond this, we aim to dig deeper to determine when and why certain approachesperform well (or badly) on both in-domain and out-of-domain data, an issue that has to date received relatively little attention.},
  url   = {https://aclanthology.org/2022.lrec-1.487}
}

@InProceedings{kim-EtAl:2022:LREC3,
  author   = {Kim, Jenia and Verkijk, Stella and Geleijn, Edwin and Leeden, Marieke van der and Meskers, Carel and Meskers, Caroline and Veen, Sabina van der and Vossen, Piek and Widdershoven, Guy},
  title   = {Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4577--4585},
  abstract = {Electronic Health Records contain a lot of information in natural language that is not expressed in the structured clinical data. Especially in the case of new diseases such as COVID-19, this information is crucial to get a better understanding of patient recovery patterns and factors that may play a role in it. However, the language in these records is very different from standard language and generic natural language processing tools cannot easily be applied out-of-the-box. In this paper, we present a fine-tuned Dutch language model specifically developed for the language in these health records that can determine the functional level of patients according to a standard coding framework from the World Health Organization. We provide evidence that our classification performs at a sufficient level to generate patient recovery patterns that can be used in the future to analyse factors that contribute to the rehabilitation of COVID-19 patients and to predict individual patient recovery of functioning.},
  url   = {https://aclanthology.org/2022.lrec-1.488}
}

@InProceedings{baimukan-bouamor-habash:2022:LREC,
  author   = {Baimukan, Nurpeiis and Bouamor, Houda and Habash, Nizar},
  title   = {Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4586--4596},

  abstract = {Arabic is a collection of dialectal variants that are historically related but significantly different. These differences can be seen across regions, countries, and even cities in the same countries. Previous work on Arabic Dialect identification has focused mainly on specific dialect levels (region, country, province, or city) using level-specific resources; and different efforts used different schemas and labels. In this paper, we present the first effort aiming at defining a standard unified three-level hierarchical schema (region-country-city) for dialectal Arabic classification. We map 29 different data sets to this unified schema, and use the common mapping to facilitate aggregating these data sets. We test the value of such aggregation by building language models and using them in dialect identification. We make our label mapping code and aggregated language models publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.489}
}

@InProceedings{lee-EtAl:2022:LREC1,
  author   = {Lee, Meisin and Soon, Lay-Ki and Siew, Eu Gene and Sugianto, Ly Fie},
  title   = {CrudeOilNews: An Annotated Crude Oil News Corpus for Event Extraction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {465--479},
  abstract = {In this paper, we present CrudeOilNews, a corpus of English Crude Oil news for event extraction. It is the first of its kind for Commodity News and serves to contribute towards resource building for economic and financial text mining. This paper describes the data collection process, the annotation methodology, and the event typology used in producing the corpus. Firstly, a seed set of 175 news articles were manually annotated, of which a subset of 25 news was used as the adjudicated reference test set for inter-annotator and system evaluation. The inter-annotator agreement was generally substantial, and annotator performance was adequate, indicating that the annotation scheme produces consistent event annotations of high quality. Subsequently, the dataset is expanded through (1) data augmentation and (2) Human-in-the-loop active learning. The resulting corpus has 425 news articles with approximately 11k events annotated. As part of the active learning process, the corpus was used to train basic event extraction models for machine labeling; the resulting models also serve as a validation or as a pilot study demonstrating the use of the corpus in machine learning purposes. The annotated corpus is made available for academic research purpose at https://github.com/meisin/CrudeOilNews-Corpus},
  url   = {https://aclanthology.org/2022.lrec-1.49}
}

@InProceedings{wertz-EtAl:2022:LREC,
  author   = {Wertz, Lukas and Mirylenka, Katsiaryna and Kuhn, Jonas and Bogojeska, Jasmina},
  title   = {Investigating Active Learning Sampling Strategies for Extreme Multi Label Text Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4597--4605},
  abstract = {Large scale, multi-label text datasets with high numbers of different classes are expensive to annotate, even more so if they deal with domain specific language. In this work, we aim to build classifiers on these datasets using Active Learning in order to reduce the labeling effort. We outline the challenges when dealing with extreme multi-label settings and show the limitations of existing Active Learning strategies by focusing on their effectiveness as well as efficiency in terms of computational cost. In addition, we present five multi-label datasets which were compiled from hierarchical classification tasks to serve as benchmarks in the context of extreme multi-label classification for future experiments. Finally, we provide insight into multi-class, multi-label evaluation and present an improved classifier architecture on top of pre-trained transformer language models.},
  url   = {https://aclanthology.org/2022.lrec-1.490}
}

@InProceedings{kutzner-laue:2022:LREC,
  author   = {Kutzner, Kristin and Laue, Ralf},
  title   = {German Light Verb Constructions in Business Process Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4606--4610},
  abstract = {We present a resource of German light verb constructions extracted from textual labels in graphical business process models. Those models depict the activities in processes in an organization in a semi-formal way. From a large range of sources, we compiled a repository of 2,301 business process models. Their textual labels (altogether 52,963 labels) were analyzed. This produced a list of 5,246 occurrences of 846 light verb constructions. We found that the light verb constructions that occur in business process models differ from light verb constructions that have been analyzed in other texts. Hence, we conclude that texts in graphical business process models represent a specific type of texts that is worth to be studied on its own. We think that our work is a step towards better automatic analysis of business process models because understanding the actual meaning of activity labels is a

prerequisite for detecting certain types of modelling problems.},
  url   = {https://aclanthology.org/2022.lrec-1.491}
}

@InProceedings{meadows-zhou-freitas:2022:LREC,
  author   = {Meadows, Jordan and Zhou, Zili and Freitas, André},
  title   = {PhysNLU: A Language Resource for Evaluating Natural Language Understanding and Explanation Coherence in Physics},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4611--4619},
  abstract = {In order for language models to aid physics research, they must first encode representations of mathematical and natural language discourse which lead to coherent explanations, with correct ordering and relevance of statements. We present a collection of datasets developed to evaluate the performance of language models in this regard, which measure capabilities with respect to sentence ordering, position, section prediction, and discourse coherence. Analysis of the data reveals the classes of arguments and sub-disciplines which are most common in physics discourse, as well as the sentence-level frequency of equations and expressions. We present baselines that demonstrate how contemporary language models are challenged by coherence related tasks in physics, even when trained on mathematical natural language objectives.},
  url   = {https://aclanthology.org/2022.lrec-1.492}
}

@InProceedings{todirascu-EtAl:2022:LREC,
  author   = {Todirascu, Amalia and Wilkens, Rodrigo and Rolin, Eva and François, Thomas and Bernhard, Delphine and Gala, Núria},
  title   = {HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4620--4630},
  abstract = {Reducing the complexity of texts by applying an Automatic Text Simplification (ATS) system has been sparking interest inthe area of Natural Language Processing (NLP) for several years and a number of methods and evaluation campaigns haveemerged targeting lexical and syntactic transformations. In recent years, several studies exploit deep learning techniques basedon very large comparable corpora. Yet the lack of large amounts of corpora (original-simplified) for French has been hinderingthe development of an ATS tool for this language. In this paper, we present our

system, which is based on a combination ofmethods relying on word embeddings for lexical simplification and rule-based strategies for syntax and discourse adaptations. We present an evaluation of the lexical, syntactic and discourse-level simplifications according to automatic and humanevaluations. We discuss the performances of our system at the lexical, syntactic, and discourse levels},
  url   = {https://aclanthology.org/2022.lrec-1.493}
}

@InProceedings{henrichsen-fuglsangengmose:2022:LREC,
  author   = {Henrichsen, Peter Juel and Fuglsang Engmose, Stine},
  title   = {AiRO - an Interactive Learning Tool for Children at Risk of Dyslexia},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4631--4636},
  abstract = {This paper presents the AiRO learning tool, which is designed for use in classrooms and homes by children at risk of developing dyslexia. The tool is based on the client-server architecture with a graphical and auditive front end (providing the interaction with the learner) and all NLP-related components located at the back end (analysing the pupil's input, deciding on the system's response, preparing speech synthesis and other feedback, logging the pupil's performance etc). AiRO software consists of independent modules for easy maintenance, e.g., upgrading the didactics or preparing AiROs for other languages. This paper also reports on our first tests 'in vivo' (November 2021) with 49 pupils (aged 6). The subjects completed 16 AiRO sessions over a four-week period. The subjects were pre- and post-tested on spelling and reading. The experimental group significantly out-performed the control group, suggesting that a new IT-supported teaching strategy may be within reach. A collection of AiRO resources (language materials, software, synthetic voice) are available as open source. At LREC, we shall present a demo of the AiRO learning tool.},
  url   = {https://aclanthology.org/2022.lrec-1.494}
}

@InProceedings{simonsen-EtAl:2022:LREC,
  author   = {Simonsen, Annika and Lamhauge, Sandra Saxov and Debess, Iben Nyholm and Henrichsen, Peter Juel},
  title   = {Creating a Basic Language Resource Kit for Faroese},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4637--4643},
  abstract = {The biggest challenges we face in developing LR and

LT for Faroese is the lack of existing resources. A few resources already exist for Faroese, but many of them are either of insufficient size and quality or are not easily accessible. Therefore, the Faroese ASR project, Ravnur, set out to make a BLARK for Faroese. The BLARK is still in the making, but many of its resources have already been produced or collected. The LR status is framed by mentioning existing LR of relevant size and quality. The specific components of the BLARK are presented as well as the working principles behind the BLARK. The BLARK will be a pillar in Faroese LR, being relatively substantial in both size, quality, and diversity. It will be open-source, inviting other small languages to use it as an inspiration to create their own BLARK. We comment on the faulty yet sprouting LT situation in the Faroe Islands. The LR and LT challenges are not solved with just a BLARK. Some initiatives are therefore proposed to better the prospects of Faroese LT. The open-source principle of the project should facilitate further development.},
  url   = {https://aclanthology.org/2022.lrec-1.495}
}

@InProceedings{ladttir-EtAl:2022:LREC,
  author   = {Óladóttir, Hulda and Arnardóttir, Þórunn and Ingason, Anton and Þorsteinsson, Vilhjálmur},
  title   = {Developing a Spell and Grammar Checker for Icelandic using an Error Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4644--4653},
  abstract = {A lack of datasets for spelling and grammatical error correction in Icelandic, along with language-specific issues, has caused a dearth of spell and grammar checking systems for the language. We present the first open-source spell and grammar checking tool for Icelandic, using an error corpus at all stages. This error corpus was in part created to aid in the development of the tool. The system is built with a rule-based tool stack comprising a tokenizer, a morphological tagger, and a parser. For token-level error annotation, tokenization rules, word lists, and a trigram model are used in error detection and correction. For sentence-level error annotation, we use specific error grammar rules in the parser as well as regex-like patterns to search syntax trees. The error corpus gives valuable insight into the errors typically made when Icelandic text is written, and guided each development phase in a test-driven manner. We assess the system’s performance with both automatic and human evaluation, using the test set in the error corpus as a reference in the automatic evaluation. The data in the error corpus development set proved useful in various ways for error detection and correction.},
  url   = {https://aclanthology.org/2022.lrec-1.496}
}

@InProceedings{suresh-EtAl:2022:LREC,
  author   = {Suresh, Abhijit and Jacobs, Jennifer and Harty, Charis and Perkoff, Margaret and Martin, James H. and Sumner, Tamara},
  title   = {The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4654--4662},
  abstract = {Transcripts of teaching episodes can be effective tools to understand discourse patterns in classroom instruction. According to most educational experts, sustained classroom discourse is a critical component of equitable, engaging, and rich learning environments for students. This paper describes the TalkMoves dataset, composed of 567 human-annotated K-12 mathematics lesson transcripts (including entire lessons or portions of lessons) derived from video recordings. The set of transcripts primarily includes in-person lessons with whole-class discussions and/or small group work, as well as some online lessons. All of the transcripts are human-transcribed, segmented by the speaker (teacher or student), and annotated at the sentence level for ten discursive moves based on accountable talk theory. In addition, the transcripts include utterance-level information in the form of dialogue act labels based on the Switchboard Dialog Act Corpus. The dataset can be used by educators, policymakers, and researchers to understand the nature of teacher and student discourse in K-12 math classrooms. Portions of this dataset have been used to develop the TalkMoves application, which provides teachers with automated, immediate, and actionable feedback about their mathematics instruction.},
  url   = {https://aclanthology.org/2022.lrec-1.497}
}

@InProceedings{gecchele-EtAl:2022:LREC,
  author   = {Gecchele, Marcello and Yamada, Hiroaki and Tokunaga, Takenobu and Sawaki, Yasuyo and Ishizuka, Mika},
  title   = {Automating Idea Unit Segmentation and Alignment for Assessing Reading Comprehension via Summary Protocol Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4663--4673},
  abstract = {In this paper, we approach summary evaluation from an applied linguistics (AL) point of view. We provide computational tools to AL researchers to simplify the process of Idea Unit (IU) segmentation. The IU is a segmentation unit that can identify chunks of information. These chunks can be compared across documents to measure the content overlap between a summary and its source text.

We propose a full revision of the annotation guidelines to allow machine implementation. The new guideline also improves the inter-annotator agreement, rising from 0.547 to 0.785 (Cohen's Kappa). We release L2WS 2021, a IU gold standard corpus composed of 40 manually annotated student summaries. We propose IUExtract; i.e. the first automatic segmentation algorithm based on the IU. The algorithm was tested over the L2WS 2021 corpus. Our results are promising, achieving a precision of 0.789 and a recall of 0.844. We tested an existing approach to IU alignment via word embeddings with the state of the art model SBERT. The recorded precision for the top 1 aligned pair of IUs was 0.375. We deemed this result insufficient for effective automatic alignment. We propose "SAT", an online tool to facilitate the collection of alignment gold standards for future training.},
  url   = {https://aclanthology.org/2022.lrec-1.498}
}

@InProceedings{singh-EtAl:2022:LREC1,
  author   = {Singh, Keshav and Inoue, Naoya and Mim, Farjana Sultana and Naito, Shoichi and Inui, Kentaro},
  title   = {IRAC: A Domain-Specific Annotated Corpus of Implicit Reasoning in Arguments},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4674--4683},
  abstract = {The task of implicit reasoning generation aims to help machines understand arguments by inferring plausible reasonings (usually implicit) between argumentative texts. While this task is easy for humans, machines still struggle to make such inferences and deduce the underlying reasoning. To solve this problem, we hypothesize that as human reasoning is guided by innate collection of domain-specific knowledge, it might be beneficial to create such a domain-specific corpus for machines. As a starting point, we create the first domain-specific resource of implicit reasonings annotated for a wide range of arguments, which can be leveraged to empower machines with better implicit reasoning generation ability. We carefully design an annotation framework to collect them on a large scale through crowdsourcing and show the feasibility of creating a such a corpus at a reasonable cost and high-quality. Our experiments indicate that models trained with domain-specific implicit reasonings significantly outperform domain-general models in both automatic and human evaluations. To facilitate further research towards implicit reasoning generation in arguments, we present an in-depth analysis of our corpus and crowdsourcing methodology, and release our materials (i.e., crowdsourcing guidelines and domain-specific resource of implicit reasonings).},
  url   = {https://aclanthology.org/2022.lrec-1.499}
}

@InProceedings{remijnse-EtAl:2022:LREC,

  author   = {Remijnse, Levi and Vossen, Piek and Fokkens, Antske and Titarsolej, Sam},
  title   = {Introducing Frege to Fillmore: A FrameNet Dataset that Captures both Sense and Reference},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {39--50},
  abstract = {This article presents the first output of the Dutch FrameNet annotation tool, which facilitates both referential- and frame annotations of language-independent corpora. On the referential level, the tool links in-text mentions to structured data, grounding the text in the real world. On the frame level, those same mentions are annotated with respect to their semantic sense. This way of annotating not only generates a rich linguistic dataset that is grounded in real-world event instances, but also guides the annotators in frame identification, resulting in high inter-annotator-agreement and consistent annotations across documents and at discourse level, exceeding traditional sentence level annotations of frame elements. Moreover, the annotation tool features a dynamic lexical lookup that increases the development of a cross-domain FrameNet lexicon.},
  url   = {https://aclanthology.org/2022.lrec-1.5}
}

@InProceedings{dehio-ostendorff-rehm:2022:LREC,
  author   = {Dehio, Niklas and Ostendorff, Malte and Rehm, Georg},
  title   = {Claim Extraction and Law Matching for COVID-19-related Legislation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {480--490},
  abstract = {To cope with the COVID-19 pandemic, many jurisdictions have introduced new or altered existing legislation. Even though these new rules are often communicated to the public in news articles, it remains challenging for laypersons to learn about what is currently allowed or forbidden since news articles typically do not reference underlying laws. We investigate an automated approach to extract legal claims from news articles and to match the claims with their corresponding applicable laws. We examine the feasibility of the two tasks concerning claims about COVID-19-related laws from Berlin, Germany. For both tasks, we create and make publicly available the data sets and report the results of initial experiments. We obtain promising results with Transformer-based models that achieve 46.7 F1 for claim extraction and 91.4 F1 for law matching, albeit with some conceptual limitations.

Furthermore, we discuss challenges of current machine learning approaches for legal language processing and their ability for complex legal reasoning tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.50}
}

@InProceedings{linke-EtAl:2022:LREC,
  author   = {Linke, Julian and Garner, Philip N. and Kubin, Gernot and Schuppler, Barbara},
  title   = {Conversational Speech Recognition Needs Data? Experiments with Austrian German},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4684--4691},
  abstract = {Conversational speech represents one of the most complex of automatic speech recognition (ASR) tasks owing to the high inter-speaker variation in both pronunciation and conversational dynamics. Such complexity is particularly sensitive to low-resourced (LR) scenarios. Recent developments in self-supervision have allowed such scenarios to take advantage of large amounts of otherwise unrelated data. In this study, we characterise an (LR) Austrian German conversational task. We begin with a non-pre-trained baseline and show that fine-tuning of a model pre-trained using self-supervision leads to improvements consistent with those in the literature; this extends to cases where a lexicon and language model are included. We also show that the advantage of pre-training indeed arises from the larger database rather than the self-supervision. Further, by use of a leave-one-conversation out technique, we demonstrate that robustness problems remain with respect to inter-speaker and inter-conversation variation. This serves to guide where future research might best be focused in light of the current state-of-the-art.},
  url   = {https://aclanthology.org/2022.lrec-1.500}
}

@InProceedings{liyanage-buscaldi-nazarenko:2022:LREC,
  author   = {Liyanage, Vijini and Buscaldi, Davide and Nazarenko, Adeline},
  title   = {A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4692--4700},
  abstract = {Automatic text generation based on neural language models has achieved performance levels that make the generated text almost indistinguishable from those written by humans. Despite the

value that text generation can have in various applications, it can also be employed for malicious tasks. The diffusion of such practices represent a threat to the quality of academic publishing. To address these problems, we propose in this paper two datasets comprised of artificially generated research content: a completely synthetic dataset and a partial text substitution dataset. In the first case, the content is completely generated by the GPT-2 model after a short prompt extracted from original papers. The partial or hybrid dataset is created by replacing several sentences of abstracts with sentences that are generated by the Arxiv-NLP model. We evaluate the quality of the datasets comparing the generated texts to aligned original texts using fluency metrics such as BLEU and ROUGE. The more natural the artificial texts seem, the more difficult they are to detect and the better is the benchmark. We also evaluate the difficulty of the task of distinguishing original from generated text by using state-of-the-art classification models.},
  url   = {https://aclanthology.org/2022.lrec-1.501}
}

@InProceedings{lauriola-small-moschitti:2022:LREC,
  author   = {Lauriola, Ivano and Small, Kevin and Moschitti, Alessandro},
  title   = {Building a Dataset for Automatically Learning to Detect Questions Requiring Clarification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4701--4707},
  abstract = {Question Answering (QA) systems aim to return correct and concise answers in response to user questions. QA research generally assumes all questions are intelligible and unambiguous, which is unrealistic in practice as questions frequently encountered by virtual assistants are ambiguous or noisy. In this work, we propose to make QA systems more robust via the following two-step process: (1) classify if the input question is intelligible and (2) for such questions with contextual ambiguity, return a clarification question. We describe a new open-domain clarification corpus containing user questions sampled from Quora, which is useful for building machine learning approaches to solving these tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.502}
}

@InProceedings{kolb-EtAl:2022:LREC,
  author   = {Kolb, Thomas and Katharina, Sekanina and Kern, Bettina Manuela Johanna and Neidhardt, Julia and Wissik, Tanja and Baumann, Andreas},
  title   = {The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4708--4716},
  abstract = {This paper introduces the Austrian German sentiment dictionary ALPIN to account for the lack of resources for dictionary-based sentiment analysis in this specific variety of German, which is characterized by lexical idiosyncrasies that also affect word sentiment. The proposed language resource is based on Austrian news media in the field of politics, an austriacism list based on different resources and a posting data set based on a popular Austrian news media. Different resources are used to increase the diversity of the resulting language resource. Extensive crowd-sourcing is performed followed by evaluation and automatic conversion into sentiment scores. We show that crowd-sourcing enables the creation of a sentiment dictionary for the Austrian German domain. Additionally, the different parts of the sentiment dictionary are evaluated to show their impact on the resulting resource. Furthermore, the proposed dictionary is utilized in a web application and available for future research and free to use for anyone.},
  url   = {https://aclanthology.org/2022.lrec-1.503}
}

@InProceedings{nghiem-EtAl:2022:LREC,
  author   = {Nghiem, Minh-Quoc and Baylis, Paul and Freitas, André and Ananiadou, Sophia},
  title   = {Text Classification and Prediction in the Legal Domain},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4717--4722},
  abstract = {We present a case study on the application of text classification and legal judgment prediction for flight compensation. We combine transformer-based classification models to classify responses from airlines and incorporate text data with other data types to predict a legal claim being successful. Our experimental evaluations show that our models achieve consistent and significant improvements over baselines and even outperformed human prediction when predicting a claim being successful. These models were integrated into an existing claim management system, providing substantial productivity gains for handling the case lifecycle, currently supporting several thousands of monthly processes.},
  url   = {https://aclanthology.org/2022.lrec-1.504}
}

@InProceedings{luecking-EtAl:2022:LREC,
  author   = {Luecking, Andy and Stoeckel, Manuel and Abrami, Giuseppe and Mehler, Alexander},

  title   = {I still have Time(s): Extending HeidelTime for German Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4723--4728},
  abstract = {HeidelTime is one of the most widespread and successful tools for detecting temporal expressions in texts. Since HeidelTime’s pattern matching system is based on regular expression, it can be extended in a convenient way. We present such an extension for the German resources of HeidelTime: HeidelTimeExt. The extension has been brought about by means of observing false negatives within real world texts and various time banks. The gain in coverage is 2.7 \% or 8.5 \%, depending on the admitted degree of potential overgeneralization. We describe the development of HeidelTimeExt, its evaluation on text samples from various genres, and share some linguistic observations. HeidelTimeExt can be obtained from https://github.com/texttechnologylab/heideltime.},
  url   = {https://aclanthology.org/2022.lrec-1.505}
}

@InProceedings{hrica-EtAl:2022:LREC,
  author   = {Hržica, Gordana and Liebeskind, Chaya and Despot, Kristina Š. and Dontcheva-Navratilova, Olga and Kamandulytė-Merfeldienė, Laura and Košutar, Sara and Kramarić, Matea and Valūnaitė Oleškevičienė, Giedrė},
  title   = {Morphological Complexity of Children Narratives in Eight Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4729--4738},
  abstract = {The aim of this study was to compare the morphological complexity in a corpus representing the language production of younger and older children across different languages. The language samples were taken from the Frog Story subcorpus of the CHILDES corpora, which comprises oral narratives collected by various researchers between 1990 and 2005. We extracted narratives by typically developing, monolingual, middle-class children. Additionally, samples of Lithuanian language, collected according to the same principles, were added. The corpus comprises 249 narratives evenly distributed across eight languages: Croatian, English, French, German, Italian, Lithuanian, Russian and Spanish. Two subcorpora were formed for each language: a younger children corpus and an older children corpus. Four measures of morphological complexity were calculated for each subcorpus: Bane, Kolmogorov, Word entropy and Relative entropy of word structure. The results showed that younger children corpora had lower morphological

complexity than older children corpora for all four measures for Spanish and Russian. Reversed results were obtained for English and French, and the results for the remaining four languages showed variation. Relative entropy of word structure proved to be indicative of age differences. Word entropy and relative entropy of word structure show potential to demonstrate typological differences.},
  url   = {https://aclanthology.org/2022.lrec-1.506}
}

@InProceedings{bucur-EtAl:2022:LREC,
  author   = {Bucur, Ana-Maria and Chitez, Madalina and Muresan, Valentina and Dinca, Andreea and Rogobete, Roxana},
  title   = {EXPRES Corpus for A Field-specific Automated Exploratory Study of L2 English Expert Scientific Writing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4739--4746},
  abstract = {Field Specific Expert Scientific Writing in English as a Lingua Franca is essential for the effective research networking and dissemination worldwide. Extracting the linguistic profile of the research articles written in L2 English can help young researchers and expert scholars in various disciplines adapt to the scientific writing norms of their communities of practice. In this exploratory study, we present and test an automated linguistic assessment model that includes features relevant for the cross-disciplinary second language framework: Text Complexity Analysis features, such as Syntactic and Lexical Complexity, and Field Specific Academic Word Lists. We analyse how these features vary across four disciplinary fields (Economics, IT, Linguistics and Political Science) in a corpus of L2-English Expert Scientific Writing, part of the EXPRES corpus (Corpus of Expert Writing in Romanian and English). The variation in field specific writing is also analysed in groups of linguistic features extracted from the higher visibility (Hv) versus lower visibility (Lv) journals. After applying lexical sophistication, lexical variation and syntactic complexity formulae, significant differences between disciplines were identified, mainly that research articles from Lv journals have higher lexical complexity, but lower syntactic complexity than articles from Hv journals; while academic vocabulary proved to have discipline specific variation.},
  url   = {https://aclanthology.org/2022.lrec-1.507}
}

@InProceedings{mullick-EtAl:2022:LREC2,
  author   = {Mullick, Ankan and Nandy, Abhilash and Kapadnis, Manav and Patnaik, Sohan and R, Raghav and Kar, Roshni},
  title   = {An Evaluation Framework for Legal Document Summarization},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4747--4753},
  abstract = {A law practitioner has to go through numerous lengthy legal case proceedings for their practices of various categories, such as land dispute, corruption, etc. Hence, it is important to summarize these documents, and ensure that summaries contain phrases with intent matching the category of the case. To the best of our knowledge, there is no evaluation metric that evaluates a summary based on its intent. We propose an automated intent-based summarization metric, which shows a better agreement with human evaluation as compared to other automated metrics like BLEU, ROUGE-L etc. in terms of human satisfaction. We also curate a dataset by annotating intent phrases in legal documents, and show a proof of concept as to how this system can be automated.},
  url   = {https://aclanthology.org/2022.lrec-1.508}
}

@InProceedings{charmet-EtAl:2022:LREC,
  author   = {Charmet, Thibault and Cherichi, Inès and Allain, Matthieu and Czerwinska, Urszula and Fouret, Amaury and Sagot, Benoît and Bawden, Rachel},
  title   = {Complex Labelling and Similarity Prediction in Legal Texts: Automatic Analysis of France's Court of Cassation Rulings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4754--4766},
  abstract = {Detecting divergences in the applications of the law (where the same legal text is applied differently by two rulings) is an important task. It is the mission of the French Cour de Cassation. The first step in the detection of divergences is to detect similar cases, which is currently done manually by experts. They rely on summarised versions of the rulings (syntheses and keyword sequences), which are currently produced manually and are not available for all rulings. There is also a high degree of variability in the keyword choices and the level of granularity used. In this article, we therefore aim to provide automatic tools to facilitate the search for similar rulings. We do this by (i) providing automatic keyword sequence generation models, which can be used to improve the coverage of the analysis, and (ii) providing measures of similarity based on the available texts and augmented with predicted keyword sequences. Our experiments show that the predictions improve correlations of automatically obtained similarities against our specially colelcted human judgments of similarity.},
  url   = {https://aclanthology.org/2022.lrec-1.509}
}

@InProceedings{ali-EtAl:2022:LREC,
  author   = {Ali, Basit and Pawar, Sachin and Palshikar, Girish and Singh, Rituraj},
  title   = {Constructing A Dataset of Support and Attack Relations in Legal Arguments in Court Judgements using Linguistic Rules},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {491--500},
  abstract = {Argumentation mining is a growing area of research and has several interesting practical applications of mining legal arguments. Support and Attack relations are the backbone of any legal argument. However, there is no publicly available dataset of these relations in the context of legal arguments expressed in court judgements. In this paper, we focus on automatically constructing such a dataset of Support and Attack relations between sentences in a court judgment with reasonable accuracy. We propose three sets of rules based on linguistic knowledge and distant supervision to identify such relations from Indian Supreme Court judgments. The first rule set is based on multiple discourse connectors, the second rule set is based on common semantic structures between argumentative sentences in a close neighbourhood, and the third rule set uses the information about the source of the argument. We also explore a BERT-based sentence pair classification model which is trained on this dataset. We release the dataset of 20506 sentence pairs - 10746 Support (precision 77.3\%) and 9760 Attack (precision 65.8\%). We believe that this dataset and the ideas explored in designing the linguistic rules and will boost the argumentation mining research for legal arguments.},
  url   = {https://aclanthology.org/2022.lrec-1.51}
}

@InProceedings{tleubayev-EtAl:2022:LREC,
  author   = {Tleubayev, Bolat and Zhexenova, Zhanel and Koishybay, Kenessary and Sandygulova, Anara},
  title   = {Cyrillic-MNIST: a Cyrillic Version of the MNIST Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4767--4773},
  abstract = {This paper presents a new handwritten dataset, Cyrillic-MNIST, a Cyrillic version of the MNIST dataset, comprising of 121,234 samples of 42 Cyrillic letters. The performance of Cyrillic-MNIST is evaluated using standard deep learning approaches and is compared to the Extended MNIST (EMNIST) dataset. The dataset

is available at https://github.com/bolattleubayev/cmnist},
  url   = {https://aclanthology.org/2022.lrec-1.510}
}

@InProceedings{barry-EtAl:2022:LREC,
  author   = {Barry, James and Wagner, Joachim and Cassidy, Lauren and Cowap, Alan and Lynn, Teresa and Walsh, Abigail and Ó Meachair, Mícheál J. and Foster, Jennifer},
  title   = {gaBERT — an Irish Language Model},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4774--4788},
  abstract = {The BERT family of neural language models have become highly popular due to their ability to provide sequences of text with rich context-sensitive token encodings which are able to generalise well to many NLP tasks. We introduce gaBERT, a monolingual BERT model for the Irish language. We compare our gaBERT model to multilingual BERT and the monolingual Irish WikiBERT, and we show that gaBERT provides better representations for a downstream parsing task. We also show how different filtering criteria, vocabulary size and the choice of subword tokenisation model affect downstream performance. We compare the results of fine-tuning a gaBERT model with an mBERT model for the task of identifying verbal multiword expressions, and show that the fine-tuned gaBERT model also performs better at this task. We release gaBERT and related code to the community.},
  url   = {https://aclanthology.org/2022.lrec-1.511}
}

@InProceedings{heeringa-EtAl:2022:LREC,
  author   = {Heeringa, Wilbert and Bouma, Gosse and Hofman, Martha and Brouwer, Jelle and Drenth, Eduard and Wijffels, Jan and Van de Velde, Hans},
  title   = {PoS Tagging, Lemmatization and Dependency Parsing of West Frisian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4789--4798},
  abstract = {We present a lemmatizer/PoS tagger/dependency parser for West Frisian using a corpus of 44,714 words in 3,126 sentences that were annotated according to the guidelines of Universal Dependencies version 2. PoS tags were assigned to words by using a Dutch PoS tagger that was applied to a Dutch word-by-word translation, or to sentences of a Dutch parallel text. Best results were obtained when using word-by-word translations that were created by using the previous version of the Frisian translation program

Oersetter. Morphologic and syntactic annotations were generated on the basis of a Dutch word-by-word translation as well. The performance of the lemmatizer/tagger/annotator when it was trained using default parameters was compared to the performance that was obtained when using the parameter values that were used for training the LassySmall UD 2.5 corpus. We study the effects of different hyperparameter settings on the accuracy of the annotation pipeline. The Frisian lemmatizer/PoS tagger/dependency parser is released as a web app and as a web service.},
  url   = {https://aclanthology.org/2022.lrec-1.512}
}

@InProceedings{plakidis-rehm:2022:LREC,
  author   = {Plakidis, Melina and Rehm, Georg},
  title   = {A Dataset of Offensive German Language Tweets Annotated for Speech Acts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4799--4807},
  abstract = {We present a dataset consisting of German offensive and non-offensive tweets, annotated for speech acts. These 600 tweets are a subset of the dataset by Struß et al. (2019) and comprises three levels of annotation, i.e., six coarse-grained speech acts, 23 fine-grained speech acts and 14 different sentence types. Furthermore, we provide an evaluation in both qualitative and quantitative terms. The dataset is made publicly available under a CC-BY-4.0 license.},
  url   = {https://aclanthology.org/2022.lrec-1.513}
}

@InProceedings{krielke-EtAl:2022:LREC,
  author   = {Krielke, Marie-Pauline and Talamo, Luigi and Fawzi, Mahmoud and Knappen, Jörg},
  title   = {Tracing Syntactic Change in the Scientific Genre: Two Universal Dependency-parsed Diachronic Corpora of Scientific English and German},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4808--4816},
  abstract = {We present two comparable diachronic corpora of scientific English and German from the Late Modern Period (17th c.–19th c.) annotated with Universal Dependencies. We describe several steps of data pre-processing and evaluate the resulting parsing accuracy showing how our pre-processing steps significantly improve output quality. As a sanity check for the representativity of our data, we conduct a case study comparing previously gained insights

on grammatical change in the scientific genre with our data. Our results reflect the often reported trend of English scientific discourse towards heavy noun phrases and a simplification of the sentence structure (Halliday, 1988; Halliday and Martin, 1993; Biber and Gray, 2011; Biber and Gray, 2016). We also show that this trend applies to German scientific discourse as well. The presented corpora are valuable resources suitable for the contrastive analysis of syntactic diachronic change in the scientific genre between 1650 and 1900. The presented pre-processing procedures and their evaluations are applicable to other languages and can be useful for a variety of Natural Language Processing tasks such as syntactic parsing.},
  url   = {https://aclanthology.org/2022.lrec-1.514}
}

@InProceedings{morgadodacosta-bond-winder:2022:LREC,
  author   = {Morgado da Costa, Luís and Bond, Francis and Winder, Roger V. P.},
  title   = {The Tembusu Treebank: An English Learner Treebank},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4817--4826},
  abstract = {This paper reports on the creation and development of the Tembusu Learner Treebank – an open treebank created from the NTU Corpus of Learner English, unique for incorporating mal-rules in the annotation of ungrammatical sentences. It describes the motivation and development of the treebank, as well as its exploitation to build a new parse-ranking model for the English Resource Grammar, designed to help improve the parse selection of ungrammatical sentences and diagnose these sentences through mal-rules. The corpus contains 25,000 sentences, of which 4,900 are treebanked. The paper concludes with an evaluation experiment that shows the usefulness of this new treebank in the tasks of grammatical error detection and diagnosis.},
  url   = {https://aclanthology.org/2022.lrec-1.515}
}

@InProceedings{ksen-EtAl:2022:LREC,
  author   = {Kåsen, Andre and Hagen, Kristin and Nøklestad, Anders and Priestly, Joel and Solberg, Per Erik and Haug, Dag Trygve Truslew},
  title   = {The Norwegian Dialect Corpus Treebank},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4827--4832},
  abstract = {This paper presents the NDC Treebank of spoken

Norwegian dialects in the Bokmål variety of Norwegian. It consists of dialect recordings made between 2006 and 2012 which have been digitised, segmented, transcribed and subsequently annotated with morphological and syntactic analysis. The nature of the spoken data gives rise to various challenges both in segmentation and annotation. We follow earlier efforts for Norwegian, in particular the LIA Treebank of spoken dialects transcribed in the Nynorsk variety of Norwegian, in the annotation principles to ensure interusability of the resources. We have developed a spoken language parser on the basis of the annotated material and report on its accuracy both on a test set across the dialects and by holding out single dialects.},
  url   = {https://aclanthology.org/2022.lrec-1.516}
}

@InProceedings{bladier-EtAl:2022:LREC,
  author   = {Bladier, Tatiana and Evang, Kilian and Generalova, Valeria and Ghane, Zahra and Kallmeyer, Laura and Möllemann, Robin and Moors, Natalia and Osswald, Rainer and Petitjean, Simon},
  title   = {RRGparbank: A Parallel Role and Reference Grammar Treebank},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4833--4841},
  abstract = {This paper describes the first release of RRGparbank, a multilingual parallel treebank for Role and Reference Grammar (RRG) containing annotations of George Orwell’s novel 1984 and its translations. The release comprises the entire novel for English and a constructionally diverse and highly parallel sample (“seed”) for German, French and Russian. The paper gives an overview of annotation decisions that have been taken and describes the adopted treebanking methodology. Finally, as a possible application, a multilingual parser is trained on the treebank data. RRGparbank is one of the first resources to apply RRG to large amounts of real-world data. Furthermore, it enables comparative and typological corpus studies in RRG. And, finally, it creates new possibilities of data-driven NLP applications based on RRG.},
  url   = {https://aclanthology.org/2022.lrec-1.517}
}

@InProceedings{chiarcos-fth-ionov:2022:LREC2,
  author   = {Chiarcos, Christian and Fäth, Christian and Ionov, Maxim},
  title   = {Unifying Morphology Resources with OntoLex-Morph. A Case Study in German},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {4842--4850},
abstract = {The OntoLex vocabulary has become a widely used community standard for machine-readable lexical resources on the web. The primary motivation to use OntoLex in favor of tool- or application-specific formalisms is to facilitate interoperability and information integration across different resources. One of its extension that is currently being developed is a module for representing morphology, OntoLex-Morph. In this paper, we show how OntoLex-Morph can be used for the encoding and integration of different types of morphological resources on a unified basis. With German as the example, we demonstrate it for (a) a full-form dictionary with inflection information (Unimorph), (b) a dictionary of base forms and their derivations (UDer), (c) a dictionary of compounds (from GermaNet), and (d) lexicon and inflection rules of a finite-state parser/generator (SMOR/Morphisto). These data are converted to OntoLex-Morph, their linguistic information is consolidated and corresponding lexical entries are linked with each other.},
url = {https://aclanthology.org/2022.lrec-1.518}
}

@InProceedings{asakura-miyao-aizawa:2022:LREC,
author = {Asakura, Takuto and Miyao, Yusuke and Aizawa, Akiko},
title = {Building Dataset for Grounding of Formulae — Annotating Coreference Relations Among Math Identifiers},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {4851--4858},
abstract = {Grounding the meaning of each symbol in math formulae is important for automated understanding of scientific documents. Generally speaking, the meanings of math symbols are not necessarily constant, and the same symbol is used in multiple meanings. Therefore, coreference relations between symbols need to be identified for grounding, and the task has aspects of both description alignment and coreference analysis. In this study, we annotated 15 papers selected from arXiv.org with the grounding information. In total, 12,352 occurrences of math identifiers in these papers were annotated, and all coreference relations between them were made explicit in each paper. The constructed dataset shows that regardless of the ambiguity of symbols in math formulae, coreference relations can be labeled with a high inter-annotator agreement. The constructed dataset enables us to achieve automation of formula grounding, and in turn, make deeper use of the knowledge in scientific documents using techniques such as math information extraction. The built grounding dataset is available at https://sigmathling.kwarc.info/resources/grounding- dataset/.},
url = {https://aclanthology.org/2022.lrec-1.519}

}

@InProceedings{paccosi-palmeroaprosio:2022:LREC,
  author   = {Paccosi, Teresa and Palmero Aprosio, Alessio},
  title   = {KIND: an Italian Multi-Domain Dataset for Named Entity Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {501--507},
  abstract = {In this paper we present KIND, an Italian dataset for Named-entity recognition. It contains more than one million tokens with annotation covering three classes: person, location, and organization. The dataset (around 600K tokens) mostly contains manual gold annotations in three different domains (news, literature, and political discourses) and a semi-automatically annotated part. The multi-domain feature is the main strength of the present work, offering a resource which covers different styles and language uses, as well as the largest Italian NER dataset with manual gold annotations. It represents an important resource for the training of NER systems in Italian. Texts and annotations are freely downloadable from the Github repository.},
  url   = {https://aclanthology.org/2022.lrec-1.52}
}

@InProceedings{nedoluzhko-EtAl:2022:LREC2,
  author   = {Nedoluzhko, Anna and Novák, Michal and Popel, Martin and Žabokrtský, Zdeněk and Zeldes, Amir and Zeman, Daniel},
  title   = {CorefUD 1.0: Coreference Meets Universal Dependencies},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4859--4872},
  abstract = {Recent advances in standardization for annotated language resources have led to successful large scale efforts, such as the Universal Dependencies (UD) project for multilingual syntactically annotated data. By comparison, the important task of coreference resolution, which clusters multiple mentions of entities in a text, has yet to be standardized in terms of data formats or annotation guidelines. In this paper we present CorefUD, a multilingual collection of corpora and a standardized format for coreference resolution, compatible with morphosyntactic annotations in the UD framework and including facilities for related tasks such as named entity recognition, which forms a first step in the direction of convergence for coreference resolution across languages.},

  url   = {https://aclanthology.org/2022.lrec-1.520}
}

@InProceedings{yu-EtAl:2022:LREC2,
  author   = {Yu, Juntao and Khosla, Sopan and Moosavi, Nafise Sadat and Paun, Silviu and Pradhan, Sameer and Poesio, Massimo},
  title   = {The Universal Anaphora Scorer},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4873--4883},
  abstract = {The aim of the Universal Anaphora initiative is to push forward the state of the art in anaphora and anaphora resolution by expanding the aspects of anaphoric interpretation which are or can be reliably annotated in anaphoric corpora, producing unified standards to annotate and encode these annotations, deliver datasets encoded according to these standards, and developing methods for evaluating models carrying out this type of interpretation. Such expansion of the scope of anaphora resolution requires a comparable expansion of the scope of the scorers used to evaluate this work. In this paper, we introduce an extended version of the Reference Coreference Scorer (Pradhan et al., 2014) that can be used to evaluate the extended range of anaphoric interpretation included in the current Universal Anaphora proposal. The UA scorer supports the evaluation of identity anaphora resolution and of bridging reference resolution, for which scorers already existed but not integrated in a single package. It also supports the evaluation of split antecedent anaphora and discourse deixis, for which no tools existed. The proposed approach to the evaluation of split antecedent anaphora is entirely novel; the proposed approach to the evaluation of discourse deixis leverages the encoding of discourse deixis proposed in Universal Anaphora to enable the use for discourse deixis of the same metrics already used for identity anaphora. The scorer was tested in the recent CODI-CRAC 2021 Shared Task on Anaphora Resolution in Dialogues.},
  url   = {https://aclanthology.org/2022.lrec-1.521}
}

@InProceedings{zhukova-hamborg-gipp:2022:LREC,
  author   = {Zhukova, Anastasia and Hamborg, Felix and Gipp, Bela},
  title   = {Towards Evaluation of Cross-document Coreference Resolution Models Using Datasets with Diverse Annotation Schemes},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4884--4893},

  abstract = {Established cross-document coreference resolution (CDCR) datasets contain event-centric coreference chains of events and entities with identity relations. These datasets establish strict definitions of the coreference relations across related tests but typically ignore anaphora with more vague context-dependent loose coreference relations. In this paper, we qualitatively and quantitatively compare the annotation schemes of ECB+, a CDCR dataset with identity coreference relations, and NewsWCL50, a CDCR dataset with a mix of loose context-dependent and strict coreference relations. We propose a phrasing diversity metric (PD) that encounters for the diversity of full phrases unlike the previously proposed metrics and allows to evaluate lexical diversity of the CDCR datasets in a higher precision. The analysis shows that coreference chains of NewsWCL50 are more lexically diverse than those of ECB+ but annotating of NewsWCL50 leads to the lower inter-coder reliability. We discuss the different tasks that both CDCR datasets create for the CDCR models, i.e., lexical disambiguation and lexical diversity. Finally, to ensure generalizability of the CDCR models, we propose a direction for CDCR evaluation that combines CDCR datasets with multiple annotation schemes that focus of various properties of the coreference chains.},
  url   = {https://aclanthology.org/2022.lrec-1.522}
}

@InProceedings{bhattarai-granmo-jiao:2022:LREC2,
  author   = {Bhattarai, Bimal and Granmo, Ole-Christoffer and Jiao, Lei},
  title   = {Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4894--4903},
  abstract = {The proliferation of fake news, i.e., news intentionally spread for misinformation, poses a threat to individuals and society. Despite various fact-checking websites such as PolitiFact, robust detection techniques are required to deal with the increase in fake news. Several deep learning models show promising results for fake news classification, however, their black-box nature makes it difficult to explain their classification decisions and quality-assure the models. We here address this problem by proposing a novel interpretable fake news detection framework based on the recently introduced Tsetlin Machine (TM). In brief, we utilize the conjunctive clauses of the TM to capture lexical and semantic properties of both true and fake news text. Further, we use clause ensembles to calculate the credibility of fake news. For evaluation, we conduct experiments on two publicly available datasets, PolitiFact and GossipCop, and demonstrate that the TM framework significantly outperforms previously published baselines by at least 5\% in terms of accuracy, with the added benefit of an interpretable logic-based representation. In addition,

our approach provides a higher F1-score than BERT and XLNet, however, we obtain slightly lower accuracy. We finally present a case study on our model's explainability, demonstrating how it decomposes into meaningful words and their negations.},
  url   = {https://aclanthology.org/2022.lrec-1.523}
}

@InProceedings{hatab-sabty-abdennadher:2022:LREC,
  author   = {Hatab, Ali L. and Sabty, Caroline and Abdennadher, Slim},
  title   = {Enhancing Deep Learning with Embedded Features for Arabic Named Entity Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4904--4912},
  abstract = {The introduction of word embedding models has remarkably changed many Natural Language Processing tasks. Word embeddings can automatically capture the semantics of words and other hidden features. Nonetheless, the Arabic language is highly complex, which results in the loss of important information. This paper uses Madamira, an external knowledge source, to generate additional word features. We evaluate the utility of adding these features to conventional word and character embeddings to perform the Named Entity Recognition (NER) task on Modern Standard Arabic (MSA). Our NER model is implemented using Bidirectional Long Short Term Memory and Conditional Random Fields (BiLSTM-CRF). We add morphological and syntactical features to different word embeddings to train the model. The added features improve the performance by different values depending on the used embedding model. The best performance is achieved by using Bert embeddings. Moreover, our best model outperforms the previous systems to the best of our knowledge.},
  url   = {https://aclanthology.org/2022.lrec-1.524}
}

@InProceedings{vakulenko-kiesel-frbe:2022:LREC,
  author   = {Vakulenko, Svitlana and Kiesel, Johannes and Fröbe, Maik},
  title   = {SCAI-QReCC Shared Task on Conversational Question Answering},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4913--4922},
  abstract = {Search-Oriented Conversational AI (SCAI) is an established venue that regularly puts a spotlight upon the recent work advancing the field of conversational search. SCAI’21 was

organised as an independent online event and featured a shared task on conversational question answering, on which this paper reports. The shared task featured three subtasks that correspond to three steps in conversational question answering: question rewriting, passage retrieval, and answer generation. This report discusses each subtask, but emphasizes the answer generation subtask as it attracted the most attention from the participants and we identified evaluation of answer correctness in the conversational settings as a major challenge and acurrent research gap. Alongside the automatic evaluation, we conducted two crowdsourcing experiments to collect annotations for answer plausibility and faithfulness. As a result of this shared task, the original conversational QA dataset used for evaluation was further extended with alternative correct answers produced by the participant systems.},
  url   = {https://aclanthology.org/2022.lrec-1.525}
}

@InProceedings{raring-ostendorff-rehm:2022:LREC,
  author   = {Raring, Michael and Ostendorff, Malte and Rehm, Georg},
  title   = {Semantic Relations between Text Segments for Semantic Storytelling: Annotation Tool - Dataset - Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4923--4932},
  abstract = {Semantic Storytelling describes the goal to automatically and semi-automatically generate stories based on extracted, processed, classified and annotated information from large content resources. Essential is the automated processing of text segments extracted from different content resources by identifying the relevance of a text segment to a topic and its semantic relation to other text segments. In this paper we present an approach to create an automatic classifier for semantic relations between extracted text segments from different news articles. We devise custom annotation guidelines based on various discourse structure theories and annotate a dataset of 2,501 sentence pairs extracted from 2,638 Wikinews articles. For the annotation, we developed a dedicated annotation tool. Based on the constructed dataset, we perform initial experiments with Transformer language models that are trained for the automatic classification of semantic relations. Our results with promising high accuracy scores suggest the validity and applicability of our approach for future Semantic Storytelling solutions.},
  url   = {https://aclanthology.org/2022.lrec-1.526}
}

@InProceedings{dhar-bisazza-vannoord:2022:LREC,
  author   = {Dhar, Prajit and Bisazza, Arianna and van Noord, Gertjan},
  title   = {Evaluating Pre-training Objectives for Low-Resource

Translation into Morphologically Rich Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4933--4943},
  abstract = {The scarcity of parallel data is a major limitation for Neural Machine Translation (NMT) systems, in particular for translation into morphologically rich languages (MRLs). An important way to overcome the lack of parallel data is to leverage target monolingual data, which is typically more abundant and easier to collect. We evaluate a number of techniques to achieve this, ranging from back-translation to random token masking, on the challenging task of translating English into four typologically diverse MRLs, under low-resource settings. Additionally, we introduce Inflection Pre-Training (or PT-Inflect), a novel pre-training objective whereby the NMT system is pre-trained on the task of re-inflecting lemmatized target sentences before being trained on standard source-to-target language translation. We conduct our evaluation on four typologically diverse target MRLs, and find that PT-Inflect surpasses NMT systems trained only on parallel data. While PT-Inflect is outperformed by back-translation overall, combining the two techniques leads to gains in some of the evaluated language pairs.},
  url   = {https://aclanthology.org/2022.lrec-1.527}
}

@InProceedings{bhattacharyya-EtAl:2022:LREC,
  author   = {Bhattacharyya, Abhidip and Mauceri, Cecilia and Palmer, Martha and Heckman, Christoffer},
  title   = {Aligning Images and Text with Semantic Role Labels for Fine-Grained Cross-Modal Understanding},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4944--4954},
  abstract = {As vision processing and natural language processing continue to advance, there is increasing interest in multimodal applications, such as image retrieval, caption generation, and human-robot interaction. These tasks require close alignment between the information in the images and text. In this paper, we present a new multimodal dataset that combines state of the art semantic annotation for language with the bounding boxes of corresponding images. This richer multimodal labeling supports cross-modal inference for applications in which such alignment is useful. Our semantic representations, developed in the natural language processing community, abstract away from the surface structure of the sentence, focusing on specific actions and the roles of their participants, a level that is equally relevant to images. We then

utilize these representations in the form of semantic role labels in the captions and the images and demonstrate improvements in standard tasks such as image retrieval. The potential contributions of these additional labels is evaluated using a role-aware retrieval system based on graph convolutional and recurrent neural networks. The addition of semantic roles into this system provides a significant increase in capability and greater flexibility for these tasks, and could be extended to state-of-the-art techniques relying on transformers with larger amounts of annotated data.},
  url   = {https://aclanthology.org/2022.lrec-1.528}
}

@InProceedings{bertinleme-EtAl:2022:LREC,
  author   = {Bertin-Lemée, Elise and Braffort, Annelies and Challant, Camille and Danet, Claire and Dauriac, Boris and Filhol, Michael and Martinod, Emmanuella and Segouat, Jérémie},
  title   = {Rosetta-LSF: an Aligned Corpus of French Sign Language and French for Text-to-Sign Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4955--4962},
  abstract = {This article presents a new French Sign Language (LSF) corpus called "Rosetta-LSF". It was created to support future studies on the automatic translation of written French into LSF, rendered through the animation of a virtual signer. An overview of the field highlights the importance of a quality representation of LSF. In order to obtain quality animations understandable by signers, it must surpass the simple "gloss transcription" of the LSF lexical units to use in the discourse. To achieve this, we designed a corpus composed of four types of aligned data, and evaluated its usability. These are: news headlines in French, translations of these headlines into LSF in the form of videos showing animations of a virtual signer, gloss annotations of the "traditional" type---although including additional information on the context in which each gestural unit is performed as well as their potential for adaptation to another context---and AZee representations of the videos, i.e. formal expressions capturing the necessary and sufficient linguistic information. This article describes this data, exhibiting an example from the corpus. It is available online for public research.},
  url   = {https://aclanthology.org/2022.lrec-1.529}
}

@InProceedings{mikhalkova-khlyupin:2022:LREC,
  author   = {Mikhalkova, Elena and Khlyupin, Alexander A.},
  title   = {Russian Jeopardy! Data Set for Question-Answering Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {508--514},
  abstract = {Question answering (QA) is one of the most common NLP tasks that relates to named entity recognition, fact extraction, semantic search and some other fields. In industry, it is much valued in chat-bots and corporate information systems. It is also a challenging task that attracted the attention of a very general audience at the quiz show Jeopardy! In this article we describe a Jeopardy!-like Russian QA data set collected from the official Russian quiz database Ch-g-k. The data set includes 379,284 quiz-like questions with 29,375 from the Russian analogue of Jeopardy! (Own Game). We observe its linguistic features and the related QA-task. We conclude about perspectives of a QA challenge based on the collected data set.},
  url   = {https://aclanthology.org/2022.lrec-1.53}
}

@InProceedings{fomicheva-EtAl:2022:LREC,
  author   = {Fomicheva, Marina and Sun, Shuo and Fonseca, Erick and Zerva, Chrysoula and Blain, Frédéric and Chaudhary, Vishrav and Guzmán, Francisco and Lopatina, Nina and Specia, Lucia and Martins, André F. T.},
  title   = {MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4963--4974},
  abstract = {We present MLQE-PE, a new dataset for Machine Translation (MT) Quality Estimation (QE) and Automatic Post-Editing (APE). The dataset contains annotations for eleven language pairs, including both high- and low-resource languages. Specifically, it is annotated for translation quality with human labels for up to 10,000 translations per language pair in the following formats: sentence-level direct assessments and post-editing effort, and word-level binary good/bad labels. Apart from the quality-related scores, each source-translation sentence pair is accompanied by the corresponding post-edited sentence, as well as titles of the articles where the sentences were extracted from, and information on the neural MT models used to translate the text. We provide a thorough description of the data collection and annotation process as well as an analysis of the annotation distribution for each language pair. We also report the performance of baseline systems trained on the MLQE-PE dataset. The dataset is freely available and has already been used for several WMT shared tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.530}
}

@InProceedings{moon-EtAl:2022:LREC2,

  author   = {Moon, Sangwhan and Cho, Won Ik and Han, Hye Joo and Okazaki, Naoaki and Kim, Nam Soo},
  title   = {OpenKorPOS: Democratizing Korean Tokenization with Voting-Based Open Corpus Annotation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4975--4983},
  abstract = {Korean is a language with complex morphology that uses spaces at larger-than-word boundaries, unlike other East-Asian languages. While morpheme-based text generation can provide significant semantic advantages compared to commonly used character-level approaches, Korean morphological analyzers only provide a sequence of morpheme-level tokens, losing information in the tokenization process. Two crucial issues are the loss of spacing information and subcharacter level morpheme normalization, both of which make the tokenization result challenging to reconstruct the original input string, deterring the application to generative tasks. As this problem originates from the conventional scheme used when creating a POS tagging corpus, we propose an improvement to the existing scheme, which makes it friendlier to generative tasks. On top of that, we suggest a fully-automatic annotation of a corpus by leveraging public analyzers. We vote the surface and POS from the outcome and fill the sequence with the selected morphemes, yielding tokenization with a decent quality that incorporates space information. Our scheme is verified via an evaluation done on an external corpus, and subsequently, it is adapted to Korean Wikipedia to construct an open, permissive resource. We compare morphological analyzer performance trained on our corpus with existing methods, then perform an extrinsic evaluation on a downstream task.},
  url   = {https://aclanthology.org/2022.lrec-1.531}
}

@InProceedings{korre-pavlopoulos:2022:LREC,
  author   = {Korre, Katerina and Pavlopoulos, John},
  title   = {Enriching Grammatical Error Correction Resources for Modern Greek},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4984--4991},
  abstract = {Grammatical Error Correction (GEC), a task of Natural Language Processing (NLP), is challenging for underepresented languages. This issue is most prominent in languages other than English. This paper addresses the issue of data and system sparsity for GEC purposes in the modern Greek Language. Following the most popular current approaches in GEC, we develop and test an MT5 multilingual text-to-text transformer for Greek. To our knowledge

this the first attempt to create a fully-fledged GEC model for Greek. Our evaluation shows that our system reaches up to 52.63\% F0.5 score on part of the Greek Native Corpus (GNC), which is 16\% below the winning system of the BEA-19 shared task on English GEC. In addition, we provide an extended version of the Greek Learner Corpus (GLC), on which our model reaches up to 22.76\% F0.5. Previous versions did not include corrections with the annotations which hindered the potential development of efficient GEC systems. For that reason we provide a new set of corrections. This new dataset facilitates an exploration of the generalisation abilities and robustness of our system, given that the assessment is conducted on learner data while the training on native data.},
  url   = {https://aclanthology.org/2022.lrec-1.532}
}

@InProceedings{mortensen-EtAl:2022:LREC,
  author   = {Mortensen, David R. and Zhang, Xinyu and Cui, Chenxuan and Zhang, Katherine},
  title   = {A Hmong Corpus with Elaborate Expression Annotations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {4992--5000},
  abstract = {This paper describes the first publicly available corpus of Hmong, a minority language of China, Vietnam, Laos, Thailand, and various countries in Europe and the Americas. The corpus has been scraped from a long-running Usenet newsgroup called soc.culture.hmong and consists of approximately 12 million tokens. This corpus (called SCH) is also the first substantial corpus to be annotated for elaborate expressions, a kind of four-part coordinate construction that is common and important in the languages of mainland Southeast Asia. We show that word embeddings trained on SCH can benefit tasks in Hmong (solving analogies) and that a model trained on it can label previously unseen elaborate expressions, in context, with an F1 of 90.79 (precision: 87.36, recall: 94.52). [ISO 639-3: mww, hmj]},
  url   = {https://aclanthology.org/2022.lrec-1.533}
}

@InProceedings{bernhard-ruizfabo:2022:LREC,
  author   = {Bernhard, Delphine and Ruiz Fabo, Pablo},
  title   = {ELAL: An Emotion Lexicon for the Analysis of Alsatian Theatre Plays},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5001--5010},

  abstract = {In this work, we present a novel and manually corrected emotion lexicon for the Alsatian dialects, including graphical variants of Alsatian lexical items. These High German dialects are spoken in the North-East of France. They are used mainly orally, and thus lack a stable and consensual spelling convention. There has nevertheless been a continuous literary production since the middle of the 17th century and, in particular, theatre plays. A large sample of Alsatian theatre plays is currently being encoded according to the Text Encoding Initiative (TEI) Guidelines. The emotion lexicon will be used to perform automatic emotion analysis in this corpus of theatre plays. We used a graph-based approach to deriving emotion scores and translations, relying only on bilingual lexicons, cognates and spelling variants. The source lexicons for emotion scores are the NRC Valence Arousal and Dominance and NRC Emotion Intensity lexicons.},
  url   = {https://aclanthology.org/2022.lrec-1.534}
}

@InProceedings{pugh-EtAl:2022:LREC,
  author   = {Pugh, Robert and Huerta Mendez, Marivel and Sasaki, Mitsuya and Tyers, Francis},
  title   = {Universal Dependencies for Western Sierra Puebla Nahuatl},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5011--5020},
  abstract = {We present a morpho-syntactically-annotated corpus of Western Sierra Puebla Nahuatl that conforms to the annotation guidelines of the Universal Dependencies project. We describe the sources of the texts that make up the corpus, the annotation process, and important annotation decisions made throughout the development of the corpus. As the first indigenous language of Mexico to be added to the Universal Dependencies project, this corpus offers a good opportunity to test and more clearly define annotation guidelines for the Meso-american linguistic area, spontaneous and elicited spoken data, and code-switching.},
  url   = {https://aclanthology.org/2022.lrec-1.535}
}

@InProceedings{baker-molla:2022:LREC,
  author   = {Baker, Gregory and Molla, Diego},
  title   = {The Construction and Evaluation of the LEAFTOP Dataset of Automatically Extracted Nouns in 1480 Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5021--5028},

  abstract = {The LEAFTOP (language extracted automatically from thousands of passages) dataset consists of nouns that appear in multiple places in the four gospels of the New Testament. We use a naive approach – probabilistic inference – to identify likely translations in 1480 other languages. We evaluate this process and find that it provides lexiconaries with accuracy from 42\% (Korafe) to 99\% (Runyankole), averaging 72\% correct across evaluated languages. The process translates up to 161 distinct lemmas from Koine Greek (average 159). We identify nouns which appear to be easy and hard to translate, language families where this technique works, and future possible improvements and extensions. The claims to novelty are: the use of a Koine Greek New Testament as the source language; using a fully-annotated manually-created grammatically parse of the source text; a custom scraper for texts in the target languages; a new metric for language similarity; a novel strategy for evaluation on low-resource languages.},
  url   = {https://aclanthology.org/2022.lrec-1.536}
}

@InProceedings{zevallos-camacho-melgarejo:2022:LREC,
  author   = {Zevallos, Rodolfo and Camacho, Luis and Melgarejo, Nelsi},
  title   = {Huqariq: A Multilingual Speech Corpus of Native Languages of Peru forSpeech Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5029--5034},
  abstract = {The Huqariq corpus is a multilingual collection of speech from native Peruvian languages. The transcribed corpus is intended for the research and development of speech technologies to preserve endangered languages in Peru. Huqariq is primarily designed for the development of automatic speech recognition, language identification and text-to-speech tools. In order to achieve corpus collection sustainably, we employs the crowdsourcing methodology. Huqariq includes four native languages of Peru, and it is expected that by the year 2022, it can reach up to 20 native languages out of the 48 native languages in Peru. The corpus has 220 hours of transcribed audio recorded by more than 500 volunteers, making it the largest speech corpus for native languages in Peru. In order to verify the quality of the corpus, we present speech recognition experiments using 220 hours of fully transcribed audio.},
  url   = {https://aclanthology.org/2022.lrec-1.537}
}

@InProceedings{vanesch-EtAl:2022:LREC,
  author   = {van Esch, Daan and Lucassen, Tamar and Ruder, Sebastian and Caswell, Isaac and Rivera, Clara},
  title   = {Writing System and Speaker Metadata for 2,800+ Language Varieties},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5035--5046},
  abstract = {We describe an open-source dataset providing metadata for about 2,800 language varieties used in the world today. Specifically, the dataset provides the attested writing system(s) for each of these 2,800+ varieties, as well as an estimated speaker count for each variety. This dataset was developed through internal research and has been used for analyses around language technologies. This is the largest publicly-available, machine-readable resource with writing system and speaker information for the world's languages. We analyze the distribution of languages and writing systems in our data and compare it to their representation in current NLP. We hope the availability of this data will catalyze research in under-represented languages.},
  url   = {https://aclanthology.org/2022.lrec-1.538}
}

@InProceedings{hagemeijer-EtAl:2022:LREC,
  author   = {Hagemeijer, Tjerk and Mendes, Amália and Gonçalves, Rita and Cornejo, Catarina and Madureira, Raquel and Généreux, Michel},
  title   = {The PALMA Corpora of African Varieties of Portuguese},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5047--5053},
  abstract = {We present three new corpora of urban varieties of Portuguese spoken in Angola, Mozambique, and São Tomé and Príncipe, where Portuguese is increasingly being spoken as first and second language in different multilingual settings. Given the scarcity of linguistic resources available for the African varieties of Portuguese, these corpora provide new, contemporary data for the study of each variety and for comparative research on African, Brazilian and European varieties, hereby improving our understanding of processes of language variation and change in postcolonial societies. The corpora consist of transcribed spoken data, complemented by a rich set of metadata describing the setting of the audio recordings and sociolinguistic information about the speakers. They are annotated with POS and lemma information and made available on the CQPweb platform, which allows for sophisticated data searches. The corpora are already being used for comparative research on constructions in the domain of possession and location involving the argument structure of intransitive, monotransitive and ditransitive verbs that select Goals, Locatives, and Recipients.},
  url   = {https://aclanthology.org/2022.lrec-1.539}
}

@InProceedings{httasch-binnig:2022:LREC,
  author   = {Hättasch, Benjamin and Binnig, Carsten},
  title   = {Know Better — A Clickbait Resolving Challenge},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {515--523},
  abstract = {In this paper, we present a new corpus of clickbait articles annotated by university students along with a corresponding shared task: clickbait articles use a headline or teaser that hides information from the reader to make them curious to open the article. We therefore propose to construct approaches that can automatically extract the relevant information from such an article, which we call clickbait resolving. We show why solving this task might be relevant for end users, and why clickbait can probably not be defeated with clickbait detection alone. Additionally, we argue that this task, although similar to question answering and some automatic summarization approaches, needs to be tackled with specialized models. We analyze the performance of some basic approaches on this task and show that models fine-tuned on our data can outperform general question answering models, while providing a systematic approach to evaluate the results. We hope that the data set and the task will help in giving users tools to counter clickbait in the future.},
  url   = {https://aclanthology.org/2022.lrec-1.54}
}

@InProceedings{maran-EtAl:2022:LREC,
  author   = {Marşan, Büşra and Yıldız, Oğuz K. and Kuzgun, Aslı and Cesur, Neslihan and Yenice, Arife B. and Sanıyar, Ezgi and Kuyrukçu, Oğuzhan and Arıcan, Bilge N. and Yıldız, Olcay Taner},
  title   = {A Learning-Based Dependency to Constituency Conversion Algorithm for the Turkish Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5054--5062},
  abstract = {This study aims to create the very first dependency-to-constituency conversion algorithm optimised for Turkish language. For this purpose, a state-of-the-art morphologic analyser and a feature-based machine learning model was used. In order to enhance the performance of the conversion algorithm, bootstrap aggregating meta-algorithm was integrated. While creating the conversation algorithm, typological properties of Turkish were carefully considered. A comprehensive and manually annotated UD-style dependency treebank was the input, and constituency trees were the

output of the conversion algorithm. A team of linguists manually annotated a set of constituency trees. These manually annotated trees were used as the gold standard to assess the performance of the algorithm. The conversion process yielded more than 8000 constituency trees whose UD-style dependency trees are also available on GitHub. In addition to its contribution to Turkish treebank resources, this study also offers a viable and easy-to-implement conversion algorithm that can be used to generate new constituency treebanks and training data for NLP resources like constituency parsers.},
  url   = {https://aclanthology.org/2022.lrec-1.540}
}

@InProceedings{mutal-EtAl:2022:LREC,
  author   = {Mutal, Jonathan David and Bouillon, Pierrette and Gerlach, Johanna and Haberkorn, Veronika},
  title   = {Standard German Subtitling of Swiss German TV content: the PASSAGE Project},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5063--5070},
  abstract = {In Switzerland, two thirds of the population speak Swiss German, a primarily spoken language with no standardised written form. It is widely used on Swiss TV, for example in news reports, interviews or talk shows, and subtitles are required for people who cannot understand this spoken language. This paper focuses on the task of automatic Standard German subtitling of spoken Swiss German, and more specifically on the translation of a normalised Swiss German speech recognition result into Standard German suitable for subtitles. Our contribution consists of a comparison of different statistical and deep learning MT systems for this task and an aligned corpus of normalised Swiss German and Standard German subtitles. Results of two evaluations, automatic and human, show that the systems succeed in improving the content, but are currently not capable of producing entirely correct Standard German.},
  url   = {https://aclanthology.org/2022.lrec-1.541}
}

@InProceedings{yadav-sitaram:2022:LREC,
  author   = {Yadav, Hemant and Sitaram, Sunayana},
  title   = {A Survey of Multilingual Models for Automatic Speech Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5071--5079},

  abstract = {Although Automatic Speech Recognition (ASR) systems have achieved human-like performance for a few languages, the majority of the world's languages do not have usable systems due to the lack of large speech datasets to train these models. Cross-lingual transfer is an attractive solution to this problem, because low-resource languages can potentially benefit from higher-resource languages either through transfer learning, or being jointly trained in the same multilingual model. The problem of cross-lingual transfer has been well studied in ASR, however, recent advances in Self Supervised Learning are opening up avenues for unlabeled speech data to be used in multilingual ASR models, which can pave the way for improved performance on low-resource languages. In this paper, we survey the state of the art in multilingual ASR models that are built with cross-lingual transfer in mind. We present best practices for building multilingual models from research across diverse languages and techniques, discuss open questions and provide recommendations for future work.},
  url   = {https://aclanthology.org/2022.lrec-1.542}
}

@InProceedings{lothritz-EtAl:2022:LREC,
  author   = {Lothritz, Cedric and Lebichot, Bertrand and Allix, Kevin and Veiber, Lisa and BISSYANDE, TEGAWENDE and Klein, Jacques and Boytsov, Andrey and Lefebvre, Clément and Goujon, Anne},
  title   = {LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5080--5089},
  abstract = {Pre-trained Language Models such as BERT have become ubiquitous in NLP where they have achieved state-of-the-art performance in most NLP tasks. While these models are readily available for English and other widely spoken languages, they remain scarce for low-resource languages such as Luxembourgish. In this paper, we present LuxemBERT, a BERT model for the Luxembourgish language that we create using the following approach: we augment the pre-training dataset by considering text data from a closely related language that we partially translate using a simple and straightforward method. We are then able to produce the LuxemBERT model, which we show to be effective for various NLP tasks: it outperforms a simple baseline built with the available Luxembourgish text data as well the multilingual mBERT model, which is currently the only option for transformer-based language models in Luxembourgish. Furthermore, we present datasets for various downstream NLP tasks that we created for this study and will make available to researchers on request.},
  url   = {https://aclanthology.org/2022.lrec-1.543}
}

@InProceedings{mohtaj-tavakkoli-asghari:2022:LREC,
  author   = {Mohtaj, Salar and Tavakkoli, Fatemeh and Asghari, Habibollah},
  title   = {PerPaDa: A Persian Paraphrase Dataset based on Implicit Crowdsourcing Data Collection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5090--5096},
  abstract = {In this paper we introduce PerPaDa, a Persian paraphrase dataset that is collected from users' input in a plagiarism detection system. As an implicit crowdsourcing experience, we have gathered a large collection of original and paraphrased sentences from Hamtajoo; a Persian plagiarism detection system, in which users try to conceal cases of text re-use in their documents by paraphrasing and re-submitting manuscripts for analysis. The compiled dataset contains 2446 instances of paraphrasing. In order to improve the overall quality of the collected data, some heuristics have been used to exclude sentences that don't meet the proposed criteria. The introduced corpus is much larger than the available datasets for the task of paraphrase identification in Persian. Moreover, there is less bias in the data compared to the similar datasets, since the users did not try some fixed predefined rules in order to generate similar texts to their original inputs.},
  url   = {https://aclanthology.org/2022.lrec-1.544}
}

@InProceedings{ezeani-EtAl:2022:LREC,
  author   = {Ezeani, Ignatius and El-Haj, Mahmoud and Morris, Jonathan and Knight, Dawn},
  title   = {Introducing the Welsh Text Summarisation Dataset and Baseline Systems},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5097--5106},
  abstract = {Welsh is an official language in Wales and is spoken by an estimated 884,300 people (29.2\% of the population of Wales). Despite this status and estimated increase in speaker numbers since the last (2011) census, Welsh remains a minority language undergoing revitalisation and promotion by Welsh Government and relevant stakeholders. As part of the effort to increase the availability of Welsh digital technology, this paper introduces the first Welsh summarisation dataset, which we provide freely for research purposes to help advance the work on Welsh summarisation. The dataset was created by Welsh speakers through manually summarising Welsh Wikipedia articles. In addition, the paper discusses the

implementation and evaluation of different summarisation systems for Welsh. The summarisation systems and results will serve as benchmarks for the development of summarisers in other minority language contexts.},
  url   = {https://aclanthology.org/2022.lrec-1.545}
}

@InProceedings{warusawithana-EtAl:2022:LREC,
  author   = {Warusawithana, Disura and Kulaweera, Nilmani and Weerasinghe, Lakshan and Karunarathne, Buddhika},
  title   = {A Systematic Approach to Derive a Refined Speech Corpus for Sinhala},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5107--5113},
  abstract = {Speech Recognition is an active research area where advances of technology have continuously driven the development of research work. However, due to the lack of adequate resources, certain languages such as Sinhala, are left to underutilize the technology. With techniques such as crowdsourcing and web scraping, several Sinhala corpora have been created and made publicly available. Despite them being large and generic, the correctness and consistency in their text data remain questionable, especially due to the lack of uniformity in the language used in the different sources of web scraped text. Addressing that requires a thorough understanding of technical and linguistic particulars pertaining to the language, which often leaves the issue unattended. We have followed a systematic approach to derive a refined corpus using a publicly available corpus for Sinhala speech recognition. In particular, we standardized the transcriptions of the corpus by removing noise in the text. Further, we applied corrections based on Sinhala linguistics. A comparative experiment shows a promising effect of the linguistic corrections by having a relative reduction of the Word-Error-Rate by 15.9\%.},
  url   = {https://aclanthology.org/2022.lrec-1.546}
}

@InProceedings{chukwuneke-EtAl:2022:LREC,
  author   = {Chukwuneke, Chiamaka and Ezeani, Ignatius and Rayson, Paul and El-Haj, Mahmoud},
  title   = {IgboBERT Models: Building and Training Transformer Models for the Igbo Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5114--5122},
  abstract = {This work presents a standard Igbo named entity

recognition (IgboNER) dataset as well as the results from training and fine-tuning state-of-the-art transformer IgboNER models. We discuss the process of our dataset creation - data collection and annotation and quality checking. We also present experimental processes involved in building an IgboBERT language model from scratch as well as fine-tuning it along with other non-Igbo pre-trained models for the downstream IgboNER task. Our results show that, although the IgboNER task benefited hugely from fine-tuning large transformer model, fine-tuning a transformer model built from scratch with comparatively little Igbo text data seems to yield quite decent results for the IgboNER task. This work will contribute immensely to IgboNLP in particular as well as the wider African and low-resource NLP efforts Keywords: Igbo, named entity recognition, BERT models, under-resourced, dataset},
  url = {https://aclanthology.org/2022.lrec-1.547}
}

@InProceedings{saulite-EtAl:2022:LREC,
  author = {Saulite, Baiba and Darģis, Roberts and Gruzitis, Normunds and Auzina, Ilze and Levāne-Petrova, Kristīne and Pretkalniņa, Lauma and Rituma, Laura and Paikens, Peteris and Znotins, Arturs and Strankale, Laine and Pokratniece, Kristīne and Poikāns, Ilmārs and Barzdins, Guntis and Skadiņa, Inguna and Baklāne, Anda and Saulespurēns, Valdis and Ziedīņš, Jānis},
  title = {Latvian National Corpora Collection – Korpuss.lv},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {5123--5129},
  abstract = {LNCC is a diverse collection of Latvian language corpora representing both written and spoken language and is useful for both linguistic research and language modelling. The collection is intended to cover diverse Latvian language use cases and all the important text types and genres (e.g. news, social media, blogs, books, scientific texts, debates, essays, etc.), taking into account both quality and size aspects. To reach this objective, LNCC is a continuous multi-institutional and multi-project effort, supported by the Digital Humanities and Language Technology communities in Latvia. LNCC includes a broad range of Latvian texts from the Latvian National Library, Culture Information Systems Centre, Latvian National News Agency, Latvian Parliament, Latvian web crawl, various Latvian publishers, and from the Latvian language corpora created by Institute of Mathematics and Computer Science and its partners, including spoken language corpora. All corpora of LNCC are re-annotated with a uniform morpho-syntactic annotation scheme which enables federated search and consistent linguistics analysis in all the LNCC corpora, as well as facilitates to select and mix various corpora for pre-training large Latvian language models like BERT and GPT.},
  url = {https://aclanthology.org/2022.lrec-1.548}
}

@InProceedings{iordache-EtAl:2022:LREC,
  author   = {Iordache, Ioan-Bogdan and Uban, Ana Sabina and Stoean, Catalin and Dinu, Liviu P.},
  title   = {Investigating the Relationship Between Romanian Financial News and Closing Prices from the Bucharest Stock Exchange},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5130--5136},
  abstract = {A new data set is gathered from a Romanian financial news website for the duration of four years. It is further refined to extract only information related to one company by selecting only paragraphs and even sentences that referred to it. The relation between the extracted sentiment scores of the texts and the stock prices from the corresponding dates is investigated using various approaches like the lexicon-based Vader tool, Financial BERT, as well as Transformer-based models. Automated translation is used, since some models could be only applied for texts in English. It is encouraging that all models, be that they are applied to Romanian or English texts, indicate a correlation between the sentiment scores and the increase or decrease of the stock closing prices.},
  url   = {https://aclanthology.org/2022.lrec-1.549}
}

@InProceedings{freitag-EtAl:2022:LREC,
  author   = {Freitag, Dayne and Cadigan, John and Sasseen, Robert and Kalmar, Paul},
  title   = {Valet: Rule-Based Information Extraction for Rapid Deployment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {524--533},
  abstract = {We present VALET, a framework for rule-based information extraction written in Python. VALET departs from legacy approaches predicated on cascading finite-state transducers, instead offering direct support for mixing heterogeneous information--lexical, orthographic, syntactic, corpus-analytic--in a succinct syntax that supports context-free idioms. We show how a handful of rules suffices to implement sophisticated matching, and describe a user interface that facilitates exploration for development and maintenance of rule sets. Arguing that rule-based information extraction is an important methodology early in the development cycle, we describe an experiment in which a VALET model is used to annotate examples for a machine learning extraction model. While learning to emulate the extraction rules, the resulting model

generalizes them, recognizing valid extraction targets the rules failed to detect.},
  url   = {https://aclanthology.org/2022.lrec-1.55}
}

@InProceedings{ivanova-washington-tyers:2022:LREC,
  author   = {Ivanova, Sardana and Washington, Jonathan and Tyers, Francis},
  title   = {A Free/Open-Source Morphological Analyser and Generator for Sakha},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5137--5142},
  abstract = {We present, to our knowledge, the first ever published morphological analyser and generator for Sakha, a marginalised language of Siberia. The transducer, developed using HFST, has coverage of solidly above 90\%, and high precision. In the development of the analyser, we have expanded linguistic knowledge about Sakha, and developed strategies for complex grammatical patterns. The transducer is already being used in downstream tasks, including computer assisted language learning applications for linguistic maintenance and computational linguistic shared tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.550}
}

@InProceedings{holden-cox-arppe:2022:LREC,
  author   = {Holden, Joshua and Cox, Christopher and Arppe, Antti},
  title   = {An Expanded Finite-State Transducer for Tsuut’ina Verbs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5143--5152},
  abstract = {This paper describes the expansion of a finite state transducer (FST) for the transitive verb system of Tsuut’ina (ISO 639-3: \texttt{srs}), a Dene (Athabaskan) language spoken in Alberta, Canada. Dene languages have unique templatic morphology, in which lexical, inflectional and derivational tiers are interlaced. Drawing on data from close to 9,000 verbal forms, the expanded model can handle a great range of common and rare argument structure types, including ditransitive and uniquely Dene object experiencer verbs. While challenges of speed remain, this expansion shows the ability of FST modelling to handle morphology of this type, and the expnded FST shows great promise for community language applications such as a morphologically informed online dictionary and word predictor, and for further FST development.This paper describes the

expansion of a finite state transducer (FST) for the transitive verb system of Tsuut’ina (ISO 639-3: \texttt{srs}), a Dene (Athabaskan) language spoken in Alberta, Canada. Dene languages have unique templatic morphology, in which lexical, inflectional and derivational tiers are interlaced. Drawing on data from over 12,000 verbs forms, the expanded model can handle a great range of common and rare argument structure types, including ditransitive and uniquely Dene object experiencer verbs. While challenges of speed remain, this expansion shows the ability of FST modelling to handle morphology of this type, and the expnded FST shows great promise for community language applications such as a morphologically informed online dictionary and word predictor, and for further FST development.},
  url   = {https://aclanthology.org/2022.lrec-1.551}
}

@InProceedings{romim-EtAl:2022:LREC,
  author   = {Romim, Nauros and Ahmed, Mosahed and Islam, Md Saiful and Sen Sharma, Arnab and Talukder, Hriteshwar and Amin, Mohammad Ruhul},
  title   = {BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5153--5162},
  abstract = {Social media platforms and online streaming services have spawned a new breed of Hate Speech (HS). Due to the massive amount of user-generated content on these sites, modern machine learning techniques are found to be feasible and cost-effective to tackle this problem. However, linguistically diverse datasets covering different social contexts in which offensive language is typically used are required to train generalizable models. In this paper, we identify the shortcomings of existing Bangla HS datasets and introduce a large manually labeled dataset BD-SHS that includes HS in different social contexts. The labeling criteria were prepared following a hierarchical annotation process, which is the first of its kind in Bangla HS to the best of our knowledge. The dataset includes more than 50,200 offensive comments crawled from online social networking sites and is at least 60\% larger than any existing Bangla HS datasets. We present the benchmark result of our dataset by training different NLP models resulting in the best one achieving an F1-score of 91.0\%. In our experiments, we found that a word embedding trained exclusively using 1.47 million comments from social media and streaming sites consistently resulted in better modeling of HS detection in comparison to other pre-trained embeddings. Our dataset and all accompanying codes is publicly available at github.com/naurosromim/hate-speech-dataset-for-Bengali-social-media},
  url   = {https://aclanthology.org/2022.lrec-1.552}
}

@InProceedings{mirzapour-EtAl:2022:LREC,
  author   = {Mirzapour, Mehdi and Ragheb, Waleed and Saeedizade, Mohammad Javad and Cousot, Kevin and Jacquenet, Helene and Carbon, Lawrence and Lafourcade, Mathieu},
  title   = {Introducing RezoJDM16k: a French KnowledgeGraph DataSet for Link Prediction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5163--5169},
  abstract = {Knowledge graphs applications, in industry and academia, motivate substantial research directions towards large-scale information extraction from various types of resources. Nowadays, most of the available knowledge graphs are either in English or multilingual. In this paper, we introduce RezoJDM16k, a French knowledge graph dataset based on RezoJDM. With 16k nodes, 832k triplets, and 53 relation types, RezoJDM16k can be employed in many NLP downstream tasks for the French language such as machine translation, question-answering, and recommendation systems. Moreover, we provide strong knowledge graph embedding baselines that are used in link prediction tasks for future benchmarking. Compared to the state-of-the-art English knowledge graph datasets used in link prediction, RezoJDM16k shows a similar promising predictive behavior.},
  url   = {https://aclanthology.org/2022.lrec-1.553}
}

@InProceedings{blache-EtAl:2022:LREC,
  author   = {Blache, Philippe and Antoine, Salomé and De Jong, Dorina and Huttner, Lena-Marie and Kerr, Emilia and Legou, Thierry and Maës, Eliot and François, Clément},
  title   = {The Badalona Corpus - An Audio, Video and Neuro-Physiological Conversational Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5170--5177},
  abstract = {We present in this paper the first natural conversation corpus recorded with all modalities and neuro-physiological signals. 5 dyads (10 participants) have been recorded three times, during three sessions (30mns each) with 4 days interval. During each session, audio and video are captured as well as the neural signal (EEG with Emotiv-EPOC) and the electro-physiological one (with Empatica-E4). This resource original in several respects. Technically, it is the first one gathering all these types of data in a natural conversation situation. Moreover, the recording of the same dyads at different periods opens the door

to new longitudinal investigations such as the evolution of interlocutors' alignment during the time. The paper situates this new type of resources with in the literature, presents the experimental setup and describes different annotations enriching the corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.554}
}

@InProceedings{asahara:2022:LREC,
  author   = {Asahara, Masayuki},
  title   = {Reading Time and Vocabulary Rating in the Japanese Language: Large-Scale Japanese Reading Time Data Collection Using Crowdsourcing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5178--5187},
  abstract = {This study examines how differences in human vocabulary affect reading time. Specifically, we assumed vocabulary to be the random effect of research participants when applying a generalized linear mixed model to the ratings of participants in the word familiarity survey. Thereafter, we asked the participants to take part in a self-paced reading task to collect their reading times. Through fixed effect of vocabulary when applying a generalized linear mixed model to reading time, we clarified the tendency that vocabulary differences give to reading time.},
  url   = {https://aclanthology.org/2022.lrec-1.555}
}

@InProceedings{marton-sayeed:2022:LREC,
  author   = {Marton, Yuval and Sayeed, Asad},
  title   = {Thematic Fit Bits: Annotation Quality and Quantity Interplay for Event Participant Representation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5188--5197},
  abstract = {Modeling thematic fit (a verb-argument compositional semantics task) currently requires a very large burden of labeled data. We take a linguistically machine-annotated large corpus and replace corpus layers with output from higher-quality, more modern taggers. We compare the old and new corpus versions' impact on a verb-argument fit modeling task, using a high-performing neural approach. We discover that higher annotation quality dramatically reduces our data requirement while demonstrating better supervised predicate-argument classification. But in applying the model to psycholinguistic tasks outside the training objective, we see clear gains at scale, but only in one of two thematic fit estimation

tasks, and no clear gains on the other. We also see that quality improves with training size, but perhaps plateauing or even declining in one task. Last, we tested the effect of role set size. All this suggests that the quality/quantity interplay is not all you need. We replicate previous studies while modifying certain role representation details and set a new state-of-the-art in event modeling, using a fraction of the data. We make the new corpus version public.},
  url   = {https://aclanthology.org/2022.lrec-1.556}
}

@InProceedings{cabiddu-EtAl:2022:LREC,
  author   = {Cabiddu, Francesco and Bott, Lewis and Jones, Gary and Gambi, Chiara},
  title   = {ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5198--5205},
  abstract = {Language acquisition research has benefitted from the use of annotated corpora of child-directed speech to examine key questions about how children learn and process language in real-world contexts. However, a lack of sense-annotated corpora has limited investigations of child word sense disambiguation in naturalistic contexts. In this work, we sense-tagged 53 corpora of American and English speech directed to 958 target children up to 59 months of age, comprising a large-scale sample of 15,581 utterances for 12 ambiguous words. Importantly, we carefully selected target senses that we know - from previous investigations - young children understand. As such work was part of a project focused on investigating the role of verbs in child word sense disambiguation, we additionally coded for verb instances which took a target ambiguous word as verb object. We present experimental work where we leveraged our sense-tagged corpus ChiSense-12 to examine the role of verb-event structure in child word sense disambiguation, and we outline our plan to use Transformer-based computational architectures to test hypotheses on the role of different learning mechanisms underlying children word sense disambiguation performance.},
  url   = {https://aclanthology.org/2022.lrec-1.557}
}

@InProceedings{turano-strapparava:2022:LREC,
  author   = {Turano, Beatrice and Strapparava, Carlo},
  title   = {Making People Laugh like a Pro: Analysing Humor Through Stand-Up Comedy},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5206--5211},
  abstract = {The analysis of humor using computational tools has gained popularity in the past few years, and a lot of resources have been built for this purpose. However, most of these resources focus on standalone jokes or on occasional humorous sentences during presentations. In this paper I present a new dataset, SCRIPTS, built using stand-up comedy shows transcripts: the humor that this dataset collects is inserted in a larger narrative, composed of daily events made humorous by the ability of the comedian. This different perspective on the humor problem can allow us to think and study humor in a different way and possibly to open the path to new lines of research.},
  url   = {https://aclanthology.org/2022.lrec-1.558}
}

@InProceedings{hesse-EtAl:2022:LREC,
  author   = {Hesse, Christoph and Langner, Maurice and Klabunde, Ralf and Benz, Anton},
  title   = {Testing Focus and Non-at-issue Frameworks with a Question-under-Discussion-Annotated Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5212--5219},
  abstract = {We present an annotated corpus of German driving reports for the analysis of Question-under-Discussion (QUD) based information structural distinctions. Since QUDs can hardly be defined in advance for providing a corresponding tagset, several theoretical issues arise concerning the scope and quality of the corpus and the development of an appropriate annotation tool for creating the corpus. We developed the corpus for testing the adequacy of QUD-based pragmatic frameworks of information structure. First analyses of the annotated information structures show that focus-related meaning aspects are essentially confirmed, indicating a sufficent accuracy of the annotations. Assumptions on non-at-issueness expressed by non-restrictive relative clauses made in the literature seem to be too strong, given the corpus data.},
  url   = {https://aclanthology.org/2022.lrec-1.559}
}

@InProceedings{sweers-hendrickx-strik:2022:LREC,
  author   = {Sweers, Tom and Hendrickx, Iris and Strik, Helmer},
  title   = {Negation Detection in Dutch Spoken Human-Computer Conversations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {534--542},
  abstract = {Proper recognition and interpretation of negation signals in text or communication is crucial for any form of full natural language understanding. It is also essential for computational approaches to natural language processing. In this study we focus on negation detection in Dutch spoken human-computer conversations. Since there exists no Dutch (dialogue) corpus annotated for negation we have annotated a Dutch corpus sample to evaluate our method for automatic negation detection. We use transfer learning and trained NegBERT (an existing BERT implementation used for negation detection) on English data with multilingual BERT to detect negation in Dutch dialogues. Our results show that adding in-domain training material improves the results. We show that we can detect both negation cues and scope in Dutch dialogues with high precision and recall. We provide a detailed error analysis and discuss the effects of cross-lingual and cross-domain transfer learning on automatic negation detection.},
  url   = {https://aclanthology.org/2022.lrec-1.56}
}

@InProceedings{tran-miyao:2022:LREC,
  author   = {Tran, Tu-Anh and Miyao, Yusuke},
  title   = {Development of a Multilingual CCG Treebank via Universal Dependencies Conversion},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5220--5233},
  abstract = {This paper introduces an algorithm to convert Universal Dependencies (UD) treebanks to Combinatory Categorial Grammar (CCG) treebanks. As CCG encodes almost all grammatical information into the lexicon, obtaining a high-quality CCG derivation from a dependency tree is a challenging task. Our algorithm relies on hand-crafted rules to assign categories to constituents, and a non-statistical parser to derive full CCG parses given the assigned categories. To evaluate our converted treebanks, we perform lexical, sentential, and syntactic rule coverage analysis, as well as CCG parsing experiments. Finally, we discuss how our method handles complex constructions, and propose possible future extensions.},
  url   = {https://aclanthology.org/2022.lrec-1.560}
}

@InProceedings{gagliardi-tamburini:2022:LREC,
  author   = {Gagliardi, Gloria and Tamburini, Fabio},
  title   = {The Automatic Extraction of Linguistic Biomarkers as a Viable Solution for the Early Diagnosis of Mental Disorders},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5234--5242},
  abstract = {Digital Linguistic Biomarkers extracted from spontaneous language productions proved to be very useful for the early detection of various mental disorders. This paper presents a computational pipeline for the automatic processing of oral and written texts: the tool enables the computation of a rich set of linguistic features at the acoustic, rhythmic, lexical, and morphosyntactic levels. Several applications of the instrument - for the detection of Mild Cognitive Impairments, Anorexia Nervosa, and Developmental Language Disorders - are also briefly discussed.},
  url   = {https://aclanthology.org/2022.lrec-1.561}
}

@InProceedings{chow-bond:2022:LREC,
  author   = {Chow, Siew Yeng and Bond, Francis},
  title   = {Singlish Where Got Rules One? Constructing a Computational Grammar for Singlish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5243--5250},
  abstract = {Singlish is a variety of English spoken in Singapore. In this paper, we share some of its grammar features and how they are implemented in the construction of a computational grammar of Singlish as a branch of English grammar. New rules were created and existing ones from standard English grammar of the English Resource Grammar (ERG) were changed in this branch to cater to how Singlish works. In addition, Singlish lexicon was added into the grammar together with some new lexical types. We used Head-driven Phrase Structure Grammar (HPSG) as the framework for this project of a creating a working computational grammar. As part of building the language resource, we also collected and formatted some data from the internet as part of a test suite for Singlish. Finally, the computational grammar was tested against a set of gold standard trees and compared with the standard English grammar to find out how well the grammar fares in analysing Singlish.},
  url   = {https://aclanthology.org/2022.lrec-1.562}
}

@InProceedings{villaneau-said:2022:LREC,
  author   = {Villaneau, Jeanne and SAID, Farida},
  title   = {COSMOS: Experimental and Comparative Studies of Concept Representations in Schoolchildren},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5251--5260},
  abstract = {COSMOS is a multidisciplinary research project investigating schoolchildren's beliefs and representations of specific concepts under control variables (age, gender, language spoken at home). Seven concepts are studied: {\it friend, father, mother, villain, work, television} and {\it dog}. We first present the protocol used and the data collected from a survey of 184 children in two age groups (6-7 and 9-11 years) in four schools in Brittany (France). A word-level lexical study shows that children's linguistic proficiency and lexical diversity increase with age, and we observe an interaction effect between gender and age on lexical diversity as measured with MLR (Measure of Lexical Richness). In contrast, none of the control variables affects lexical density. We also present the lemmas that schoolchildren most often associate with each concept. Generalized linear mixed-effects models reveal significant effects of age, gender, and home language on some concept-lemma associations and specific interactions between age and gender. Most of the identified effects are documented in the child development literature. To better understand the process of semantic construction in children, additional lexical analyses at the n-gram, chunk, and clause levels would be helpful. We briefly present ongoing and planned work in this direction. The COSMOS data will soon be made freely available to the scientific community.},
  url   = {https://aclanthology.org/2022.lrec-1.563}
}

@InProceedings{piccirilli-schulteimwalde:2022:LREC,
  author   = {Piccirilli, Prisca and Schulte im Walde, Sabine},
  title   = {Features of Perceived Metaphoricity on the Discourse Level: Abstractness and Emotionality},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5261--5273},
  abstract = {Research on metaphorical language has shown ties between abstractness and emotionality with regard to metaphoricity; prior work is however limited to the word and sentence levels, and up to date there is no empirical study establishing the extent to which this is also true on the discourse level. This paper explores which textual and perceptual features human annotators perceive as important for the metaphoricity of discourses and expressions, and addresses two research questions more specifically. First, is a metaphorically-perceived discourse more abstract and more emotional in comparison to a literally- perceived discourse? Second, is a metaphorical expression preceded by a more metaphorical/abstract/ emotional context than a synonymous literal alternative? We used a dataset of 1,000 corpus-extracted discourses for which crowdsourced annotators (1) provided judgements on whether they perceived the discourses as more metaphorical or more literal, and (2)

systematically listed lexical terms which triggered their decisions in (1). Our results indicate that metaphorical discourses are more emotional and to a certain extent more abstract than literal discourses. However, neither the metaphoricity nor the abstractness and emotionality of the preceding discourse seem to play a role in triggering the choice between synonymous metaphorical vs. literal expressions. Our dataset is available at https://www.ims.uni-stuttgart.de/data/discourse-met-lit.},
  url   = {https://aclanthology.org/2022.lrec-1.564}
}

@InProceedings{singh-EtAl:2022:LREC2,
  author   = {Singh, Sandhya and Roy, Prapti and Sahoo, Nihar and Mallela, Niteesh and Gupta, Himanshu and Bhattacharyya, Pushpak and Savagaonkar, Milind and Sultan, Nidhi and Ramnani, Roshni and Maitra, Anutosh and Sengupta, Shubhashis},
  title   = {Hollywood Identity Bias Dataset: A Context Oriented Bias Analysis of Movie Dialogues},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5274--5285},
  abstract = {Movies reflect society and also hold power to transform opinions. Social biases and stereotypes present in movies can cause extensive damage due to their reach. These biases are not always found to be the need of storyline but can creep in as the author’s bias. Movie production houses would prefer to ascertain that the bias present in a script is the story’s demand. Today, when deep learning models can give human-level accuracy in multiple tasks, having an AI solution to identify the biases present in the script at the writing stage can help them avoid the inconvenience of stalled release, lawsuits, etc. Since AI solutions are data intensive and there exists no domain specific data to address the problem of biases in scripts, we introduce a new dataset of movie scripts that are annotated for identity bias. The dataset contains dialogue turns annotated for (i) bias labels for seven categories, viz., gender, race/ethnicity, religion, age, occupation, LGBTQ, and other, which contains biases like body shaming, personality bias, etc. (ii) labels for sensitivity, stereotype, sentiment, emotion, emotion intensity, (iii) all labels annotated with context awareness, (iv) target groups and reason for bias labels and (v) expert-driven group-validation process for high quality annotations. We also report various baseline performances for bias identification and category detection on our dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.565}
}

@InProceedings{ahn-chodroff:2022:LREC,
  author   = {Ahn, Emily and Chodroff, Eleanor},
  title   = {VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5286--5294},
  abstract = {Cross-linguistic phonetic analysis has long been limited by data scarcity and insufficient computational resources. In the past few years, the availability of large-scale cross-linguistic spoken corpora has increased dramatically, but the data still require considerable computational power and processing for downstream phonetic analysis. To facilitate large-scale cross-linguistic phonetic research in the field, we release the VoxCommunis Corpus, which contains acoustic models, pronunciation lexicons, and word- and phone-level alignments, derived from the publicly available Mozilla Common Voice Corpus. The current release includes data from 36 languages. The corpus also contains acoustic-phonetic measurements, which currently consist of formant frequencies (F1--F4) from all vowel quartiles. Major advantages of this corpus for phonetic analysis include the number of available languages, the large amount of speech per language, as well as the fact that most language datasets have dozens to hundreds of contributing speakers. We demonstrate the utility of this corpus for downstream phonetic research in a descriptive analysis of language-specific vowel systems, as well as an analysis of "uniformity" in vowel realization across languages. The VoxCommunis Corpus is free to download and use under a CC0 license.},
  url   = {https://aclanthology.org/2022.lrec-1.566}
}

@InProceedings{peverelli-vanerp-bloemendal:2022:LREC,
  author   = {Peverelli, Andrea and van Erp, Marieke and Bloemendal, Jan},
  title   = {Tracking Textual Similarities in Neo-Latin Drama Networks},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5295--5303},
  abstract = {This paper describes the first experiments towards tracking the complex and international network of text reuse within the Early Modern (XV-XVII centuries) community of Neo-Latin humanists. Our research, conducted within the framework of the TransLatin project, aims at gaining more evidence on the topic of textual similarities and semi-conscious reuse of literary models. It consists of two experiments conveyed through two main research fields (Information Retrieval and Stylometry), as a means to a better understanding of the complex and subtle literary mechanisms underlying the drama production of Modern Age authors and their transnational network of relations. The experiments led to the

construction of networks of works and authors that fashion different patterns of similarity and models of evolution and interaction between texts.},
  url   = {https://aclanthology.org/2022.lrec-1.567}
}

@InProceedings{orasmaa-EtAl:2022:LREC,
  author   = {Orasmaa, Siim and Muischnek, Kadri and Poska, Kristjan and Edela, Anna},
  title   = {Named Entity Recognition in Estonian 19th Century Parish Court Records},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5304--5313},
  abstract = {This paper presents a new historical language resource, a corpus of Estonian Parish Court records from the years 1821-1920, annotated for named entities (NE), and reports on named entity recognition (NER) experiments using this corpus. The hand-written records have been transcribed manually via a crowdsourcing project, so the transcripts are of high quality, but the variation of language and spelling is high in these documents due to dialectal variation and the fact that there was a considerable change in Estonian spelling conventions during the time of their writing. The typology of NEs for manual annotation includes 7 categories, but the inter-annotator agreement is as good as 95.0 (mean F1-score). We experimented with fine-tuning BERT-like transfer learning approaches for NER, and found modern Estonian BERT models highly applicable, despite the difficulty of the historical material. Our best model, finetuned Est-RoBERTa, achieved microaverage F1 score of 93.6, which is comparable to state-of-the-art NER performance on the contemporary Estonian.},
  url   = {https://aclanthology.org/2022.lrec-1.568}
}

@InProceedings{eichel-lapesa-schulteimwalde:2022:LREC,
  author   = {Eichel, Annerose and Lapesa, Gabriella and Schulte im Walde, Sabine},
  title   = {Investigating Independence vs. Control: Agenda-Setting in Russian News Coverage on Social Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5314--5323},
  abstract = {Agenda-setting is a widely explored phenomenon in political science: powerful stakeholders (governments or their financial supporters) have control over the media and set their agenda: political and economical powers determine which news should

be salient. This is a clear case of targeted manipulation to divert the public attention from serious issues affecting internal politics (such as economic downturns and scandals) by flooding the media with potentially distracting information. We investigate agenda-setting in the Russian social media landscape, exploring the relation between economic indicators and mentions of foreign geopolitical entities, as well as of Russia itself. Our contributions are at three levels: at the level of the domain of the investigation, our study is the first to substructure the Russian media landscape in state-controlled vs. independent outlets in the context of strategic distraction from negative economic trends; at the level of the scope of the investigation, we involve a large set of geopolitical entities (while previous work has focused on the U.S.); at the qualitative level, our analysis of posts on Ukraine, whose relationship with Russia is of high geopolitical relevance, provides further insights into the contrast between state-controlled and independent outlets.},
  url   = {https://aclanthology.org/2022.lrec-1.569}
}

@InProceedings{cieri-EtAl:2022:LREC,
  author   = {Cieri, Christopher and Liberman, Mark and Cho, Sunghye and Strassel, Stephanie and Fiumara, James and Wright, Jonathan},
  title   = {Reflections on 30 Years of Language Resource Development and Sharing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {543--550},
  abstract = {The Linguistic Data Consortium was founded in 1992 to solve the problem that limitations in access to shareable data was impeding progress in Human Language Technology research and development. At the time, DARPA had adopted the common task research management paradigm to impose additional rigor on their programs by also providing shared objectives, data and evaluation methods. Early successes underscored the promise of this paradigm but also the need for a standing infrastructure to host and distribute the shared data. During LDC’s initial five year grant, it became clear that the demand for linguistic data could not easily be met by the existing providers and that a dedicated data center could add capacity first for data collection and shortly thereafter for annotation. The expanding purview required expansions of LDC’s technical infrastructure including systems support and software development. An open question for the center would be its role in other kinds of research beyond data development. Over its 30 years history, LDC has performed multiple roles ranging from neutral, independent data provider to multisite programs, to creator of exploratory data in tight collaboration with system developers, to research group focused on data intensive investigations.},
  url   = {https://aclanthology.org/2022.lrec-1.57}

}

@InProceedings{stymne-stman:2022:LREC,
  author   = {Stymne, Sara and Östman, Carin},
  title   = {SLäNDa version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5324--5333},
  abstract = {In this paper, we describe version 2.0 of the SLäNDa corpus. SLäNDa, the Swedish Literary corpus of Narrative and Dialogue, now contains excerpts from 19 novels, written between 1809--1940. The main focus of the SLäNDa corpus is to distinguish between direct speech and the main narrative. In order to isolate the narrative, we also annotate everything else which does not belong to the narrative, such as thoughts, quotations, and letters. SLäNDa version 2.0 has a slightly updated annotation scheme from version 1.0. In addition, we added new texts from eleven authors and performed quality control on the previous version. We are specifically interested in different ways of marking speech segments, such as quotation marks, dashes, or no marking at all. To allow a detailed evaluation of this aspect, we added dedicated test sets to SLäNDa for these different types of speech marking. In a pilot experiment, we explore the impact of typographic speech marking by using these test sets, as well as artificially stripping the training data of speech markers.},
  url   = {https://aclanthology.org/2022.lrec-1.570}
}

@InProceedings{degraaf-EtAl:2022:LREC,
  author   = {de Graaf, Evelien and Stopponi, Silvia and Bos, Jasper K. and Peels-Matthey, Saskia and Nissim, Malvina},
  title   = {AGILe: The First Lemmatizer for Ancient Greek Inscriptions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5334--5344},
  abstract = {To facilitate corpus searches by classicists as well as to reduce data sparsity when training models, we focus on the automatic lemmatization of ancient Greek inscriptions, which have not received as much attention in this sense as literary text data has. We show that existing lemmatizers for ancient Greek, trained on literary data, are not performant on epigraphic data, due to major language differences between the two types of texts. We thus train the first inscription-specific lemmatizer achieving above 80\% accuracy, and make both the models and the lemmatized data available

to the community. We also provide a detailed error analysis highlighting peculiarities of inscriptions which again highlights the importance of a lemmatizer dedicated to inscriptions.},
  url   = {https://aclanthology.org/2022.lrec-1.571}
}

@InProceedings{schauffler-EtAl:2022:LREC,
  author   = {Schauffler, Nadja and Bernhart, Toni and Blessing, Andre and Eschenbach, Gunilla and Gärtner, Markus and Jung, Kerstin and Kinder, Anna and Koch, Julia and Richter, Sandra and Viehhauser, Gabriel and Vu, Ngoc Thang and Wesemann, Lorenz and Kuhn, Jonas},
  title   = {»textklang« – Towards a Multi-Modal Exploration Platform for German Poetry},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5345--5355},
  abstract = {We present the steps taken towards an exploration platform for a multi-modal corpus of German lyric poetry from the Romantic era developed in the project »textklang«. This interdisciplinary project develops a mixed-methods approach for the systematic investigation of the relationship between written text (here lyric poetry) and its potential and actual sonic realisation (in recitations, musical performances etc.). The multi-modal »textklang« platform will be designed to technically and analytically combine three modalities: the poetic text, the audio signal of a recorded recitation and, at a later stage, music scores of a musical setting of a poem. The methodological workflow will enable scholars to develop hypotheses about the relationship between textual form and sonic/prosodic realisation based on theoretical considerations, text interpretation and evidence from recorded recitations. The full workflow will support hypothesis testing either through systematic corpus analysis alone or with addtional contrastive perception experiments. For the experimental track, researchers will be enabled to manipulate prosodic parameters in (re-)synthesised variants of the original recordings. The focus of this paper is on the design of the base corpus and on tools for systematic exploration – placing special emphasis on our response to challenges stemming from multi-modality and the methodologically diverse interdisciplinary setup.},
  url   = {https://aclanthology.org/2022.lrec-1.572}
}

@InProceedings{nguyen-wintner:2022:LREC,
  author   = {Nguyen, Isabelle and Wintner, Shuly},
  title   = {Predicting the Proficiency Level of Nonnative Hebrew Authors},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5356--5365},
  abstract = {We present classifiers that can accurately predict the proficiency level of nonnative Hebrew learners. This is important for practical (mainly educational) applications, but the endeavor also sheds light on the features that support the classification, thereby improving our understanding of learner language in general, and transfer effects from Arabic, French, and Russian on nonnative Hebrew in particular.},
  url   = {https://aclanthology.org/2022.lrec-1.573}
}

@InProceedings{vajjala:2022:LREC,
  author   = {Vajjala, Sowmya},
  title   = {Trends, Limitations and Open Challenges in Automatic Readability Assessment Research},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5366--5377},
  abstract = {Readability assessment is the task of evaluating the reading difficulty of a given piece of text. This article takes a closer look at contemporary NLP research on developing computational models for readability assessment, identifying the common approaches used for this task, their shortcomings, and some challenges for the future. Where possible, the survey also connects computational research with insights from related work in other disciplines such as education and psychology.},
  url   = {https://aclanthology.org/2022.lrec-1.574}
}

@InProceedings{das-EtAl:2022:LREC,
  author   = {Das, Mithun and Saha, Punyajoy and Mathew, Binny and Mukherjee, Animesh},
  title   = {HateCheckHIn: Evaluating Hindi Hate Speech Detection Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5378--5387},
  abstract = {Due to the sheer volume of online hate, the AI and NLP communities have started building models to detect such hateful content. Recently, multilingual hate is a major emerging challenge for automated detection where code-mixing or more than one language have been used for conversation in social media. Typically, hate speech detection models are evaluated by measuring their performance

on the held-out test data using metrics such as accuracy and F1-score. While these metrics are useful, it becomes difficult to identify using them where the model is failing, and how to resolve it. To enable more targeted diagnostic insights of such multilingual hate speech models, we introduce a set of functionalities for the purpose of evaluation. We have been inspired to design this kind of functionalities based on real-world conversation on social media. Considering Hindi as a base language, we craft test cases for each functionality. We name our evaluation dataset HateCheckHIn. To illustrate the utility of these functionalities , we test state-of-the-art transformer based m-BERT model and the Perspective API.},
  url   = {https://aclanthology.org/2022.lrec-1.575}
}

@InProceedings{li-EtAl:2022:LREC2,
  author   = {Li, Irene and Fabbri, Alex and Kawamura, Rina and Liu, Yixin and Tang, Xiangru and tae, Jaesung and Shen, Chang and Ma, Sally and Mizutani, Tomoe and Radev, Dragomir},
  title   = {Surfer100: Generating Surveys From Web Resources, Wikipedia-style},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5388--5392},
  abstract = {Fast-developing fields such as Artificial Intelligence (AI) often outpace the efforts of encyclopedic sources such as Wikipedia, which either do not completely cover recently-introduced topics or lack such content entirely. As a result, methods for automatically producing content are valuable tools to address this information overload. We show that recent advances in pretrained language modeling can be combined for a two-stage extractive and abstractive approach for Wikipedia lead paragraph generation. We extend this approach to generate longer Wikipedia-style summaries with sections and examine how such methods struggle in this application through detailed studies with 100 reference human-collected surveys. This is the first study on utilizing web resources for long Wikipedia-style summaries to the best of our knowledge.},
  url   = {https://aclanthology.org/2022.lrec-1.576}
}

@InProceedings{jauhar-EtAl:2022:LREC,
  author   = {Jauhar, Sujay Kumar and Chandrasekaran, Nirupama and Gamon, Michael and White, Ryen},
  title   = {MS-LaTTE: A Dataset of Where and When To-do Tasks are Completed},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {5393--5403},
  abstract = {Tasks are a fundamental unit of work in the daily lives of people, who are increasingly using digital means to keep track of, organize, triage, and act on them. These digital tools -- such as task management applications -- provide a unique opportunity to study and understand tasks and their connection to the real world, and through intelligent assistance, help people be more productive. By logging signals such as text, timestamp information, and social connectivity graphs, an increasingly rich and detailed picture of how tasks are created and organized, what makes them important, and who acts on them, can be progressively developed. Yet the context around actual task completion remains fuzzy, due to the basic disconnect between actions taken in the real world and telemetry recorded in the digital world. Thus, in this paper we compile and release a novel, real-life, large-scale dataset called MS-LaTTE that captures two core aspects of the context surrounding task completion: location and time. We describe our annotation framework and conduct a number of analyses on the data that were collected, demonstrating that it captures intuitive contextual properties for common tasks. Finally, we test the dataset on the two problems of predicting spatial and temporal task co-occurrence, concluding that predictors for co-location and co-time are both learnable, with a BERT fine-tuned model outperforming several other baselines. The MS-LaTTE dataset provides an opportunity to tackle many new modeling challenges in contextual task understanding and we hope that its release will spur future research in task intelligence more broadly.},
  url   = {https://aclanthology.org/2022.lrec-1.577}
}

@InProceedings{mussakhojayeva-khassanov-varol:2022:LREC,
  author   = {Mussakhojayeva, Saida and Khassanov, Yerbolat and Varol, Huseyin Atakan},
  title   = {KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5404--5411},
  abstract = {We present an expanded version of our previously released Kazakh text-to-speech (KazakhTTS) synthesis corpus. In the new KazakhTTS2 corpus, the overall size has increased from 93 hours to 271 hours, the number of speakers has risen from two to five (three females and two males), and the topic coverage has been diversified with the help of new sources, including a book and Wikipedia articles. This corpus is necessary for building high-quality TTS systems for Kazakh, a Central Asian agglutinative language from the Turkic family, which presents several linguistic challenges. We describe the corpus construction process and provide the details of the training and evaluation procedures for the TTS

system. Our experimental results indicate that the constructed corpus is sufficient to build robust TTS models for real-world applications, with a subjective mean opinion score ranging from 3.6 to 4.2 for all the five speakers. We believe that our corpus will facilitate speech and language research for Kazakh and other Turkic languages, which are widely considered to be low-resource due to the limited availability of free linguistic data. The constructed corpus, code, and pretrained models are publicly available in our GitHub repository.},
  url   = {https://aclanthology.org/2022.lrec-1.578}
}

@InProceedings{oksanen-EtAl:2022:LREC,
  author   = {Oksanen, Joel and Majumder, Abhilash and Saunack, Kumar and Toni, Francesca and Dhondiyal, Arun},
  title   = {A Graph-Based Method for Unsupervised Knowledge Discovery from Financial Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5412--5417},
  abstract = {The need for manual review of various financial texts, such as company filings and news, presents a major bottleneck in financial analysts' work. Thus, there is great potential for the application of NLP methods, tools and resources to fulfil a genuine industrial need in finance. In this paper, we show how this potential can be fulfilled by presenting an end-to-end, fully unsupervised method for knowledge discovery from financial texts. Our method creatively integrates existing resources to construct automatically a knowledge graph of companies and related entities as well as to carry out unsupervised analysis of the resulting graph to provide quantifiable and explainable insights from the produced knowledge. The graph construction integrates entity processing and semantic expansion, before carrying out open relation extraction. We illustrate our method by calculating automatically the environmental rating for companies in the S\&P 500, based on company filings with the SEC (Securities and Exchange Commission). We then show the usefulness of our method in this setting by providing an assessment of our method's outputs with an independent MSCI source.},
  url   = {https://aclanthology.org/2022.lrec-1.579}
}

@InProceedings{mapelli-EtAl:2022:LREC,
  author   = {Mapelli, Valérie and Arranz, Victoria and Choukri, Khalid and Mazo, Hélène},
  title   = {Language Resources to Support Language Diversity - the ELRA Achievements},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {551--558},
  abstract = {This article highlights ELRA’s latest achievements in the field of Language Resources (LRs) identification, sharing and production. It also reports on ELRA's involvement in several national and international projects, as well as in the organization of events for the support of LRs and related Language Technologies, including for under-resourced languages. Over the past few years, ELRA, together with its operational agency ELDA, has continued to increase its catalogue offer of LRs, establishing worldwide partnerships for the production of various types of LRs (SMS, tweets, crawled data, MT aligned data, speech LRs, sentiment-based data, etc.). Through their consistent involvement in EU-funded projects, ELRA and ELDA have contributed to improve the access to multilingual information in the context of the pandemic, develop tools for the de-identification of texts in the legal and medical domains, support the EU eTranslation Machine Translation system, and set up a European platform providing access to both resources and services. In December 2019, ELRA co-organized the LT4All conference, whose main topics were Language Technologies for enabling linguistic diversity and multilingualism worldwide. Moreover, although LREC was cancelled in 2020, ELRA published the LREC 2020 proceedings for the Main conference and Workshops papers, and carried on its dissemination activities while targeting the new LREC edition for 2022.},
  url   = {https://aclanthology.org/2022.lrec-1.58}
}

@InProceedings{boinepelli-EtAl:2022:LREC,
  author   = {Boinepelli, Sravani and Raha, Tathagata and Abburi, Harika and Parikh, Pulkit and Chhaya, Niyati and Varma, Vasudeva},
  title   = {Leveraging Mental Health Forums for User-level Depression Detection on Social Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5418--5427},
  abstract = {The number of depression and suicide risk cases on social media platforms is ever-increasing, and the lack of depression detection mechanisms on these platforms is becoming increasingly apparent. A majority of work in this area has focused on leveraging linguistic features while dealing with small-scale datasets. However, one faces many obstacles when factoring into account the vastness and inherent imbalance of social media content. In this paper, we aim to optimize the performance of user-level depression classification to lessen the burden on computational resources. The resulting system executes in a quicker, more efficient manner, in turn making it suitable for deployment. To simulate a platform agnostic framework, we simultaneously replicate

the size and composition of social media to identify victims of depression. We systematically design a solution that categorizes post embeddings, obtained by fine-tuning transformer models such as RoBERTa, and derives user-level representations using hierarchical attention networks. We also introduce a novel mental health dataset to enhance the performance of depression categorization. We leverage accounts of depression taken from this dataset to infuse domain-specific elements into our framework. Our proposed methods outperform numerous baselines across standard metrics for the task of depression detection in text.},
  url   = {https://aclanthology.org/2022.lrec-1.580}
}

@InProceedings{danielsson-EtAl:2022:LREC,
  author   = {Danielsson, Benjamin and Santini, Marina and Lundberg, Peter and Al-Abasse, Yosef and Jonsson, Arne and Eneling, Emma and Stridsman, Magnus},
  title   = {Classifying Implant-Bearing Patients via their Medical Histories: a Pre-Study on Swedish EMRs with Semi-Supervised GanBERT},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5428--5435},
  abstract = {In this paper, we compare the performance of two BERT-based text classifiers whose task is to classify patients (more precisely, their medical histories) as having or not having implant(s) in their body. One classifier is a fully-supervised BERT classifier. The other one is a semi-supervised GAN-BERT classifier. Both models are compared against a fully-supervised SVM classifier. Since fully-supervised classification is expensive in terms of data annotation, with the experiments presented in this paper, we investigate whether we can achieve a competitive performance with a semi-supervised classifier based only on a small amount of annotated data. Results are promising and show that the semi-supervised classifier has a competitive performance with the fully-supervised classifier.},
  url   = {https://aclanthology.org/2022.lrec-1.581}
}

@InProceedings{kchaou-EtAl:2022:LREC,
  author   = {Kchaou, Saméh and boujelbane, rahma and Fsih, Emna and Hadrich-Belguith, Lamia},
  title   = {Standardisation of Dialect Comments in Social Networks in View of Sentiment Analysis : Case of Tunisian Dialect},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {5436--5443},
  abstract = {With the growing access to the internet, the spoken Arabic dialect language becomes informal languages written in social media. Most users post comments using their own dialect. This linguistic situation inhibits mutual understanding between internet users and makes difficult to use computational approaches since most Arabic resources are intended for the formal language: Modern Standard Arabic (MSA). In this paper, we present a pipeline to standardize the written texts in social networks by translating them to the standard language MSA. We fine-tun at first an identification bert-based model to select Tunisian Dialect (TD) from MSA and other dialects. Then, we learned transformer model to translate TD to MSA. The final system includes the translated TD text and the originally text written in MSA. Each of these steps was evaluated on the same test corpus. In order to test the effectiveness of the approach, we compared two opinion analysis models, the first intended for the Sentiment Analysis (SA) of dialect texts and the second for the MSA texts. We concluded that through standardization we obtain the best score.},
  url   = {https://aclanthology.org/2022.lrec-1.582}
}

@InProceedings{sosea-caragea:2022:LREC,
  author   = {Sosea, Tiberiu and Caragea, Cornelia},
  title   = {EnsyNet: A Dataset for Encouragement and Sympathy Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5444--5449},
  abstract = {More and more people turn to Online Health Communities to seek social support during their illnesses. By interacting with peers with similar medical conditions, users feel emotionally and socially supported, which in turn leads to better adherence to therapy. Current studies in Online Health Communities focus only on the presence or absence of emotional support, while the available datasets are scarce or limited in terms of size. To enable development on emotional support detection, we introduce EnsyNet, a dataset of 6,500 sentences annotated with two types of support: encouragement and sympathy. We train BERT-based classifiers on this dataset, and apply our best BERT model in two large scale experiments. The results of these experiments show that receiving encouragements or sympathy improves users' emotional state, while the lack of emotional support negatively impacts patients' emotional state.},
  url   = {https://aclanthology.org/2022.lrec-1.583}
}

@InProceedings{himoro-parejalora:2022:LREC,
  author   = {Himoro, Marcelo Yuji and Pareja-Lora, Antonio},
  title   = {Preliminary Results on the Evaluation of

Computational Tools for the Analysis of Quechua and Aymara},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5450--5459},
  abstract = {This research has focused on evaluating the existing open-source morphological analyzers for two of the most widely spoken indigenous macrolanguages in South America, namely Quechua and Aymara. Firstly, we have evaluated their performance (precision, recall and F1 score) for the individual languages for which they were developed (Cuzco Quechua and Aymara). Secondly, in order to assess how these tools handle other individual languages of the macrolanguage, we have extracted some sample text from school textbooks and educational resources. This sample text was edited in the different countries where these macrolanguages are spoken (Colombia, Ecuador, Peru, Bolivia, Chile and Argentina for Quechua; and Bolivia, Peru and Chile for Aymara), and it includes their different standardized forms (10 individual languages of Quechua and 3 of Aymara). Processing this text by means of the tools, we have (i) calculated their coverage (number of words recognized and analyzed) and (ii) studied in detail the cases for which each tool was unable to generate any output. Finally, we discuss different ways in which these tools could be optimized, either to improve their performances or, in the specific case of Quechua, to cover more individual languages of this macrolanguage in future works as well.},
  url   = {https://aclanthology.org/2022.lrec-1.584}
}

@InProceedings{arora-EtAl:2022:LREC,
  author   = {Arora, Siddhant and Hosseini, Henry and Utz, Christine and Bannihatti Kumar, Vinayshekhar and Dhellemmes, Tristan and Ravichander, Abhilasha and Story, Peter and Mangat, Jasmine and Chen, Rex and Degeling, Martin and Norton, Thomas and Hupperich, Thomas and Wilson, Shomir and Sadeh, Norman},
  title   = {A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5460--5472},
  abstract = {Over the past decade, researchers have started to explore the use of NLP to develop tools aimed at helping the public, vendors, and regulators analyze disclosures made in privacy policies. With the introduction of new privacy regulations, the language of privacy policies is also evolving, and disclosures made by the same organization are not always the same in different

languages, especially when used to communicate with users who fall under different jurisdictions. This work explores the use of language technologies to capture and analyze these differences at scale. We introduce an annotation scheme designed to capture the nuances of two new landmark privacy regulations, namely the EU’s GDPR and California’s CCPA/CPRA. We then introduce the first bilingual corpus of mobile app privacy policies consisting of 64 privacy policies in English (292K words) and 91 privacy policies in German (478K words), respectively with manual annotations for 8K and 19K fine-grained data practices. The annotations are used to develop computational methods that can automatically extract "disclosures" from privacy policies. Analysis of a subset of 59 "semi-parallel" policies reveals differences that can be attributed to different regulatory regimes, suggesting that systematic analysis of policies using automated language technologies is indeed a worthwhile endeavor.},
  url   = {https://aclanthology.org/2022.lrec-1.585}
}

@InProceedings{wang-mercer-rudzicz:2022:LREC,
  author   = {Wang, Xindi and Mercer, Robert E. and Rudzicz, Frank},
  title   = {MeSHup: Corpus for Full Text Biomedical Document Indexing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5473--5483},
  abstract = {Medical Subject Heading (MeSH) indexing refers to the problem of assigning a given biomedical document with the most relevant labels from an extremely large set of MeSH terms. Currently, the vast number of biomedical articles in the PubMed database are manually annotated by human curators, which is time consuming and costly; therefore, a computational system that can assist the indexing is highly valuable. When developing supervised MeSH indexing systems, the availability of a large-scale annotated text corpus is desirable. A publicly available, large corpus that permits robust evaluation and comparison of various systems is important to the research community. We release a large scale annotated MeSH indexing corpus, MeSHup, which contains 1,342,667 full text articles, together with the associated MeSH labels and metadata, authors and publication venues that are collected from the MEDLINE database. We train an end-to-end model that combines features from documents and their associated labels on our corpus and report the new baseline.},
  url   = {https://aclanthology.org/2022.lrec-1.586}
}

@InProceedings{gao-EtAl:2022:LREC,
  author   = {Gao, Yanjun and Dligach, Dmitriy and Miller, Timothy and Tesch, Samuel and Laffin, Ryan and Churpek,

Matthew M. and Afshar, Majid},
  title   = {Hierarchical Annotation for Building A Suite of Clinical Natural Language Processing Tasks: Progress Note Understanding},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5484--5493},
  abstract = {Applying methods in natural language processing on electronic health records (EHR) data has attracted rising interests. Existing corpus and annotation focus on modeling textual features and relation prediction. However, there are a paucity of annotated corpus built to model clinical diagnostic thinking, a processing involving text understanding, domain knowledge abstraction and reasoning. In this work, we introduce a hierarchical annotation schema with three stages to address clinical text understanding, clinical reasoning and summarization. We create an annotated corpus based on a large collection of publicly available daily progress notes, a type of EHR that is time-sensitive, problem-oriented, and well-documented by the format of Subjective, Objective, Assessment and Plan (SOAP). We also define a new suite of tasks, Progress Note Understanding, with three tasks utilizing the three annotation stages. This new suite aims at training and evaluating future NLP models for clinical text understanding, clinical knowledge representation, inference and summarization.},
  url   = {https://aclanthology.org/2022.lrec-1.587}
}

@InProceedings{nguyen-EtAl:2022:LREC,
  author   = {Nguyen, Vinh Van and Nguyen, Ha and Le, Huong Thanh and Nguyen, Thai Phuong and Bui, Tan Van and Pham, Luan Nghia and Phan, Anh Tuan and Nguyen, Cong Hoang-Minh and Tran, Viet Hong and Tran, Anh Huu},
  title   = {KC4MT: A High-Quality Corpus for Multilingual Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5494--5502},
  abstract = {The multilingual parallel corpus is an important resource for many applications of natural language processing (NLP). For machine translation, the size and quality of the training corpus mainly affects the quality of the translation models. In this work, we present the method for building high-quality multilingual parallel corpus in the news domain and for some low-resource languages, including Vietnamese, Laos, and Khmer, to improve the quality of multilingual machine translation in these areas. We also publicized this one that includes 500.000 Vietnamese-Chinese

bilingual sentence pairs; 150.000 Vietnamese-Laos bilingual sentence pairs, and 150.000 Vietnamese-Khmer bilingual sentence pairs.},
  url   = {https://aclanthology.org/2022.lrec-1.588}
}

@InProceedings{jaidka:2022:LREC,
  author   = {Jaidka, Kokil},
  title   = {Developing A Multilabel Corpus for the Quality Assessment of Online Political Talk},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5503--5510},
  abstract = {This paper motivates and presents the Twitter Deliberative Politics dataset, a corpus of political tweets labeled for its deliberative characteristics. The corpus was randomly sampled from replies to US congressmen and women. It is expected to be useful to a general community of computational linguists, political scientists, and social scientists interested in the study of online political expression, computer-mediated communication, and political deliberation. The data sampling and annotation methods are discussed and classical machine learning approaches are evaluated for their predictive performance on the different deliberative facets. The paper concludes with a discussion of future work aimed at developing dictionaries for the quality assessment of online political talk in English. The dataset and a demo dashboard are available at https://github.com/kj2013/twitter-deliberative-politics.},
  url   = {https://aclanthology.org/2022.lrec-1.589}
}

@InProceedings{kamocki-witt:2022:LREC,
  author   = {Kamocki, Pawel and Witt, Andreas},
  title   = {Ethical Issues in Language Resources and Language Technology – Tentative Categorisation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {559--563},
  abstract = {Ethical issues in Language Resources and Language Technology are often invoked, but rarely discussed. This is at least partly because little work has been done to systematize ethical issues and principles applicable in the fields of Language Resources and Language Technology. This paper provides an overview of ethical issues that arise at different stages of Language Resources and Language Technology development, from the conception phase through the construction phase to the use phase. Based on this overview, the authors propose a tentative taxonomy of ethical issues in Language

Resources and Language Technology, built around five principles: Privacy, Property, Equality, Transparency and Freedom. The authors hope that this tentative taxonomy will facilitate ethical assessment of projects in the field of Language Resources and Language Technology, and structure the discussion on ethical issues in this domain, which may eventually lead to the adoption of a universally accepted Code of Ethics of the Language Resources and Language Technology community.},
  url   = {https://aclanthology.org/2022.lrec-1.59}
}

@InProceedings{hurtado:2022:LREC,
  author   = {Hurtado, Irati},
  title   = {BILinMID: A Spanish-English Corpus of the US Midwest},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5511--5516},
  abstract = {This paper describes the Bilinguals in the Midwest (BILinMID) Corpus, a comparable text corpus of the Spanish and English spoken in the US Midwest by various types of bilinguals. Unlike other areas within the US where language contact has been widely documented (e.g., the Southwest), Spanish-English bilingualism in the Midwest has been understudied despite an increase in its Hispanic population. The BILinMID Corpus contains short stories narrated in Spanish and in English by 72 speakers representing different types of bilinguals: early simultaneous bilinguals, early sequential bilinguals, and late second language learners. All stories have been transcribed and annotated using various natural language processing tools. Additionally, a user interface has also been created to facilitate searching for specific patterns in the corpus as well as to filter out results according to specified criteria. Guidelines and procedures followed to create the corpus and the user interface are described in detail in the paper. The corpus is fully available online and it might be particularly interesting for researchers working on language variation and contact.},
  url   = {https://aclanthology.org/2022.lrec-1.590}
}

@InProceedings{rajagopal-EtAl:2022:LREC,
  author   = {Rajagopal, Dheeraj and Zhang, Xuchao and Gamon, Michael and Jauhar, Sujay Kumar and Yang, Diyi and Hovy, Eduard},
  title   = {One Document, Many Revisions: A Dataset for Classification and Description of Edit Intents},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {5517--5524},
abstract = {Document authoring involves a lengthy revision process, marked by individual edits that are frequently linked to comments. Modeling the relationship between edits and comments leads to a better understanding of document evolution, potentially benefiting applications such as content summarization, and task triaging. Prior work on understanding revisions has primarily focused on classifying edit intents, but falling short of a deeper understanding of the nature of these edits. In this paper, we present explore the challenge of describing an edit at two levels: identifying the edit intent, and describing the edit using free-form text. We begin by defining a taxonomy of general edit intents and introduce a new dataset of full revision histories of Wikipedia pages, annotated with each revision's edit intent. Using this dataset, we train a classifier that achieves a 90\% accuracy in identifying edit intent. We use this classifier to train a distantly-supervised model that generates a high-level description of a revision in free-form text. Our experimental results show that incorporating edit intent information aids in generating better edit descriptions. We establish a set of baselines for the edit description task, achieving a best score of 28 ROUGE, thus demonstrating the effectiveness of our layered approach to edit understanding.},
url = {https://aclanthology.org/2022.lrec-1.591}
}

@InProceedings{cui-EtAl:2022:LREC2,
author = {Cui, Yue and Zhu, Junhui and Yang, Liner and Fang, Xuezhi and Chen, Xiaobin and Wang, Yujie and Yang, Erhong},
title = {CTAP for Chinese:A Linguistic Complexity Feature Automatic Calculation Platform},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {5525--5538},
abstract = {The construct of linguistic complexity has been widely used in language learning research. Several text analysis tools have been created to automatically analyze linguistic complexity. However, the indexes supported by several existing Chinese text analysis tools are limited and different because of different research purposes. CTAP is an open-source linguistic complexity measurement extraction tool, which prompts any research purposes. Although it was originally developed for English, the Unstructured Information Management (UIMA) framework it used allows the integration of other languages. In this study, we integrated the Chinese component into CTAP, describing the index sets it incorporated and comparing it with three linguistic complexity tools for Chinese. The index set includes four levels of 196 linguistic

complexity indexes: character level, word level, sentence level, and discourse level. So far, CTAP has implemented automatic calculation of complexity characteristics for four languages, aiming to help linguists without NLP background study language complexity.},
  url   = {https://aclanthology.org/2022.lrec-1.592}
}

@InProceedings{pftze-EtAl:2022:LREC,
  author   = {Pfütze, Dominik and Ritz, Eva and Janda, Julius and Rietsche, Roman},
  title   = {A Corpus for Suggestion Mining of German Peer Feedback},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5539--5547},
  abstract = {Peer feedback in online education becomes increasingly important to meet the demand for feedback in large scale classes, such as e.g. Massive Open Online Courses (MOOCs). However, students are often not experts in how to write helpful feedback to their fellow students. In this paper, we introduce a corpus compiled from university students' peer feedback to be able to detect suggestions on how to improve the students' work and therefore being able to capture peer feedback helpfulness. To the best of our knowledge, this corpus is the first student peer feedback corpus in German which additionally was labelled with a new annotation scheme. The corpus consists of more than 600 written feedback (about 7,500 sentences). The utilisation of the corpus is broadly ranged from Dependency Parsing to Sentiment Analysis to Suggestion Mining, etc. We applied the latter to empirically validate the utility of the new corpus. Suggestion Mining is the extraction of sentences that contain suggestions from unstructured text. In this paper, we present a new annotation scheme to label sentences for Suggestion Mining. Two independent annotators labelled the corpus and achieved an inter-annotator agreement of 0.71. With the help of an expert arbitrator a gold standard was created. An automatic classification using BERT achieved an accuracy of 75.3\%.},
  url   = {https://aclanthology.org/2022.lrec-1.593}
}

@InProceedings{li-yu-liu:2022:LREC,
  author   = {Li, Yi and Yu, Dong and liu, pengyuan},
  title   = {CLGC: A Corpus for Chinese Literary Grace Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {5548--5556},
  abstract = {In this paper, we construct a Chinese literary grace corpus, CLGC, with 10,000 texts and more than 1.85 million tokens. Multi-level annotations are provided for each text in our corpus, including literary grace level, sentence category, and figure-of-speech type. Based on the corpus, we dig deep into the correlation between fine-grained features (semantic information, part-of-speech and figure-of-speech, etc.) and literary grace level. We also propose a new Literary Grace Evaluation (LGE) task, which aims at making a comprehensive assessment of the literary grace level according to the text. In the end, we build some classification models with machine learning algorithms (such as SVM, TextCNN) to prove the effectiveness of our features and corpus for LGE. The results of our preliminary classification experiments have achieved 79.71\% on the weighted average F1-score.},
  url   = {https://aclanthology.org/2022.lrec-1.594}
}

@InProceedings{etinolu-schweitzer:2022:LREC,
  author   = {Çetinoğlu, Özlem and Schweitzer, Antje},
  title   = {Anonymising the SAGT Speech Corpus and Treebank},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5557--5564},
  abstract = {Anonymisation, that is identifying and neutralising sensitive references, is a crucial part of dataset creation. In this paper, we describe the anonymisation process of a Turkish-German code-switching corpus, namely SAGT, which consists of speech data and a treebank that is built on its transcripts. We employed a selective pseudonymisation approach where we manually identified sensitive references to anonymise and replaced them with surrogate values on the treebank side. In addition to maintaining data privacy, our primary concerns in surrogate selection were keeping the integrity of code-switching properties, morphosyntactic annotation layers, and semantics. After the treebank anonymisation, we anonymised the speech data by mapping between the treebank sentences and audio transcripts with the help of Praat scripts. The treebank is publicly available for research purposes and the audio files can be obtained via an individual licence agreement.},
  url   = {https://aclanthology.org/2022.lrec-1.595}
}

@InProceedings{suzuki-EtAl:2022:LREC1,
  author   = {Suzuki, Daisuke and Takahashi, Yujin and Yamashita, Ikumi and Aida, Taichi and Hirasawa, Tosho and Nakatsuji, Michitaka and Mita, Masato and Komachi, Mamoru},
  title   = {Construction of a Quality Estimation Dataset for Automatic Evaluation of Japanese Grammatical Error Correction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5565--5572},
  abstract = {In grammatical error correction (GEC), automatic evaluation is considered as an important factor for research and development of GEC systems. Previous studies on automatic evaluation have shown that quality estimation models built from datasets with manual evaluation can achieve high performance in automatic evaluation of English GEC. However, quality estimation models have not yet been studied in Japanese, because there are no datasets for constructing quality estimation models. In this study, therefore, we created a quality estimation dataset with manual evaluation to build an automatic evaluation model for Japanese GEC. By building a quality estimation model using this dataset and conducting a meta-evaluation, we verified the usefulness of the quality estimation model for Japanese GEC.},
  url   = {https://aclanthology.org/2022.lrec-1.596}
}

@InProceedings{shah-EtAl:2022:LREC,
  author   = {Shah, Jui and Zhang, Dongxu and Brody, Sam and McCallum, Andrew},
  title   = {Enhanced Distant Supervision with State-Change Information for Relation Extraction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5573--5579},
  abstract = {In this work, we introduce a method for enhancing distant supervision with state-change information for relation extraction. We provide a training dataset created via this process, along with manually annotated development and test sets. We present an analysis of the curation process and data, and compare it to standard distant supervision. We demonstrate that the addition of state-change information reduces noise when used for static relation extraction, and can also be used to train a relation-extraction system that detects a change of state in relations.},
  url   = {https://aclanthology.org/2022.lrec-1.597}
}

@InProceedings{gafni-prior-wintner:2022:LREC,
  author   = {Gafni, Chen and Prior, Anat and Wintner, Shuly},
  title   = {The Hebrew Essay Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {5580--5586},
  abstract = {We present the Hebrew Essay Corpus: an annotated corpus of Hebrew language argumentative essays authored by prospective higher-education students. The corpus includes both essays by native speakers, written as part of the psychometric exam that is used to assess their future success in academic studies; and essays authored by non-native speakers, with three different native languages, that were written as part of a language aptitude test. The corpus is uniformly encoded and stored. The non-native essays were annotated with target hypotheses whose main goal is to make the texts amenable to automatic processing (morphological and syntactic analysis). The corpus is available for academic purposes upon request. We describe the corpus and the error correction and annotation schemes used in its analysis. In addition to introducing this new resource, we discuss the challenges of identifying and analyzing non-native language use in general, and propose various ways for dealing with these challenges.},
  url   = {https://aclanthology.org/2022.lrec-1.598}
}

@InProceedings{koiso-EtAl:2022:LREC,
  author   = {Koiso, Hanae and Amatani, Haruka and Den, Yasuharu and Iseki, Yuriko and Ishimoto, Yuichi and Kashino, Wakako and Kawabata, Yoshiko and Nishikawa, Ken'ya and Tanaka, Yayoi and Usuda, Yasuyuki and Watanabe, Yuka},
  title   = {Design and Evaluation of the Corpus of Everyday Japanese Conversation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5587--5594},
  abstract = {We have constructed the Corpus of Everyday Japanese Conversation (CEJC) and published it in March 2022. The CEJC is designed to contain various kinds of everyday conversations in a balanced manner to capture their diversity. The CEJC features not only audio but also video data to facilitate precise understanding of the mechanism of real-life social behavior. The publication of a large-scale corpus of everyday conversations that includes video data is a new approach. The CEJC contains 200 hours of speech, 577 conversations, about 2.4 million words, and a total of 1675 conversants. In this paper, we present an overview of the corpus, including the recording method and devices, structure of the corpus, formats of video and audio files, transcription, and annotations. We then report some results of the evaluation of the CEJC in terms of conversant and conversation attributes. We show that the CEJC includes a good balance of adult conversants in terms of gender and age, as well as a variety of conversations in terms of conversation forms, places, activities, and numbers of conversants.},
  url   = {https://aclanthology.org/2022.lrec-1.599}
}

@InProceedings{pedersen-EtAl:2022:LREC,
  author   = {Pedersen, Bolette and Sørensen, Nathalie Carmen Hau and Nimb, Sanni and Flørke, Ida and Olsen, Sussi and Troelsgård, Thomas},
  title   = {Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR Lexicon},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {51--60},
  abstract = {We present The Central Word Register for Danish (COR), which is an open source lexicon project for general AI purposes funded and initiated by the Danish Agency for Digitisation as part of an AI initiative embarked by the Danish Government in 2020. We focus here on the lexical semantic part of the project (COR-S) and describe how we – based on the existing fine-grained sense inventory from Den Danske Ordbog (DDO) – compile a more AI suitable sense granularity level of the vocabulary. A three-step methodology is applied: We establish a set of linguistic principles for defining core senses in COR-S and from there, we generate a hand-crafted gold standard of 6,000 lemmas depicting how to come from the fine-grained DDO sense to the COR inventory. Finally, we experiment with a number of language models in order to automatize the sense reduction of the rest of the lexicon. The models comprise a ruled-based model that applies our linguistic principles in terms of features, a word2vec model using cosine similarity to measure the sense proximity, and finally a deep neural BERT model fine-tuned on our annotations. The rule-based approach shows best results, in particular on adjectives, however, when focusing on the average polysemous vocabulary, the BERT model shows promising results too.},
  url   = {https://aclanthology.org/2022.lrec-1.6}
}

@InProceedings{ducel-EtAl:2022:LREC,
  author   = {Ducel, Fanny and Fort, Karën and Lejeune, Gaël and Lepage, Yves},
  title   = {Do we Name the Languages we Study? The \#BenderRule in LREC and ACL articles},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {564--573},
  abstract = {This article studies the application of the \#BenderRule in Natural Language Processing (NLP) articles according to two dimensions. Firstly, in a contrastive manner, by considering two major international conferences, LREC and ACL, and secondly, in a diachronic manner, by inspecting nearly 14,000 articles over a period of time ranging from 2000 to 2020 for LREC and from 1979 to

2020 for ACL. For this purpose, we created a corpus from LREC and ACL articles from the above-mentioned periods, from which we manually annotated nearly 1,000. We then developed two classifiers to automatically annotate the rest of the corpus. Our results show that LREC articles tend to respect the \#BenderRule (80 to 90\% of them respect it), whereas 30 to 40\% of ACL articles do not. Interestingly, over the considered periods, the results appear to be stable for the two conferences, even though a rebound in ACL 2020 could be a sign of the influence of the blog post about the \#BenderRule.},
  url   = {https://aclanthology.org/2022.lrec-1.60}
}

@InProceedings{akdemir-jeon-shibuya:2022:LREC,
  author   = {Akdemir, Arda and Jeon, Yeojoo and Shibuya, Tetsuo},
  title   = {Developing Language Resources and NLP Tools for the North Korean Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5595--5600},
  abstract = {Since the division of Korea, the two Korean languages have diverged significantly over the last 70 years. However, due to the lack of linguistic source of the North Korean language, there is no DPRK-based language model. Consequently, scholars rely on the Korean language model by utilizing South Korean linguistic data. In this paper, we first present a large-scale dataset for the North Korean language. We use the dataset to train a BERT-based language model, DPRK-BERT. Second, we annotate a subset of this dataset for the sentiment analysis task. Finally, we compare the performance of different language models for masked language modeling and sentiment analysis tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.600}
}

@InProceedings{tsuchiya-yokoi:2022:LREC,
  author   = {Tsuchiya, Masatoshi and Yokoi, Yasutaka},
  title   = {Developing a Dataset of Overridden Information in Wikipedia},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5601--5608},
  abstract = {This paper proposes a new task of detecting information override. Since all information on the Web is not updated in a timely manner, the necessity is created for information that is overridden by another information source to be discarded.

The task is formalized as a binary classification problem to determine whether a reference sentence has overridden a target sentence. In investigating this task, this paper describes a construction procedure for the dataset of overridden information by collecting sentence pairs from the difference between two versions of Wikipedia. Our developing dataset shows that the old version of Wikipedia contains much overridden information and that the detection of information override is necessary.},
  url   = {https://aclanthology.org/2022.lrec-1.601}
}

@InProceedings{consoli-EtAl:2022:LREC,
  author   = {Consoli, Bernardo and dos Santos, Henrique D. P. and Ulbrich, Ana Helena D. P. S. and Vieira, Renata and Bordini, Rafael H.},
  title   = {BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5609--5616},
  abstract = {Computational medicine research requires clinical data for training and testing purposes, so the development of datasets composed of real hospital data is of utmost importance in this field. Most such data collections are in the English language, were collected in anglophone countries, and do not reflect other clinical realities, which increases the importance of national datasets for projects that hope to positively impact public health. This paper presents a new Brazilian Clinical Dataset containing over 70,000 admissions from 10 hospitals in two Brazilian states, composed of a sum total of over 2.5 million free-text clinical notes alongside data pertaining to patient information, prescription information, and exam results. This data was collected, organized, deidentified, and is being distributed via credentialed access for the use of the research community. In the course of presenting the new dataset, this paper will explore the new dataset's structure, population, and potential benefits of using this dataset in clinical AI tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.602}
}

@InProceedings{branco-EtAl:2022:LREC,
  author   = {Branco, António and Silva, João Ricardo and Gomes, Luís and António Rodrigues, João},
  title   = {Universal Grammatical Dependencies for Portuguese with CINTIL Data, LX Processing and CLARIN support},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {5617--5626},
  abstract = {The grammatical framework for the mapping between linguistic form and meaning representation known as Universal Dependencies relies on a non-constituency syntactic analysis that is centered on the notion of grammatical relation (e.g. Subject, Object, etc.). Given its core goal of providing a common set of analysis primitives suitable to every natural language, and its practical objective of fostering their computational grammatical processing, it keeps being an active domain of research in science and technology of language. This paper presents a new collection of quality language resources for the computational processing of the Portuguese language under the Universal Dependencies framework (UD). This is an all-encompassing, publicly available open collection of mutually consistent and inter-operable scientific resources that includes reliably annotated corpora, top-performing processing tools and expert support services: a new UPOS-annotated corpus, CINTIL-UPos, with 675K tokens and a new UD treebank, CINTIL-UDep Treebank, with nearly 38K sentences; a UPOS tagger, LX-UTagger, and a UD parser, LX-UDParser, trained on these corpora, available both as local stand-alone tools and as remote web-based services; and helpdesk support ensured by the Knowledge Center for the Science and Technology of Portuguese of the CLARIN research infrastructure.},
  url   = {https://aclanthology.org/2022.lrec-1.603}
}

@InProceedings{venugopal-pramod-shekhar:2022:LREC,
  author   = {Venugopal, Gayatri and Pramod, Dhanya and Shekhar, Ravi},
  title   = {CWID-hi: A Dataset for Complex Word Identification in Hindi Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5627--5636},
  abstract = {Text simplification is a method for improving the accessibility of text by converting complex sentences into simple sentences. Multiple studies have been done to create datasets for text simplification. However, most of these datasets focus on high-resource languages only. In this work, we proposed a complex word dataset for Hindi, a language largely ignored in text simplification literature. We used various Hindi knowledge annotators for annotation to capture the annotator’s language knowledge. Our analysis shows a significant difference between native and non-native annotators’ perception of word complexity. We also built an automatic complex word classifier using a soft voting approach based on the predictions from tree-based ensemble classifiers. These models behave differently for annotations made by different categories of users, such as native and non-native speakers. Our dataset and analysis will help simplify Hindi text depending on the user’s language understanding. The dataset is available at https://

zenodo.org/record/5229160.},
  url   = {https://aclanthology.org/2022.lrec-1.604}
}

@InProceedings{rozovskaya:2022:LREC,
  author   = {Rozovskaya, Alla},
  title   = {Automatic Classification of Russian Learner Errors},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5637--5647},
  abstract = {Grammatical Error Correction systems are typically
evaluated overall, without taking into consideration performance on
individual error types because system output is not annotated with
respect to error type. We introduce a tool that automatically
classifies errors in Russian learner texts. The tool takes an edit
pair consisting of the original token(s) and the corresponding
replacement and provides a grammatical error category. Manual
evaluation of the output reveals that in more than 93\% of cases the
error categories are judged as correct or acceptable. We apply the
tool to carry out a fine-grained evaluation on the performance of
two error correction systems for Russian.},
  url   = {https://aclanthology.org/2022.lrec-1.605}
}

@InProceedings{hajnicz:2022:LREC,
  author   = {Hajnicz, Elżbieta},
  title   = {Annotation of metaphorical expressions in the Basic
Corpus of Polish Metaphors},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5648--5653},
  abstract = {This paper presents a corpus of Polish texts
annotated with metaphorical expressions. It is composed of two parts
of comparable size, selected from two subcorpora of the Polish
National Corpus: the subcorpus manually annotated on morphosyntactic
level, named entities level etc., and the Polish Coreference Corpus,
with manually annotated mentions and the coreference relations
between them, but automatically annotated on the morphosyntactic
level (only the second part is actually annotated). In the paper we
briefly outline the method for identifying metaphorical expressions
in a text, based on the MIPVU procedure. The main difference is the
stress put on novel metaphors and considering neologistic
derivatives that have metaphorical properties. The annotation
procedure is based on two notions: vehicle – a part of an expression
used metaphorically, representing a source domain and its topic – a
part referring to reality, representing a target domain. Next, we

propose several features (text form, conceptual structure, conventionality and contextuality) to classify metaphorical expressions identified in texts. Additionally, some metaphorical expressions are identified as concerning personal identity matters and classified w.r.t. their properties. Finally, we analyse and evaluate the results of the annotation.},
  url   = {https://aclanthology.org/2022.lrec-1.606}
}

@InProceedings{tian-EtAl:2022:LREC1,
  author   = {Tian, Yuanhe and Qin, Han and Xia, Fei and Song, Yan},
  title   = {ChiMST: A Chinese Medical Corpus for Word Segmentation and Medical Term Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5654--5664},
  abstract = {Chinese word segmentation (CWS) and named entity recognition (NER) are two important tasks in Chinese natural language processing. To achieve good model performance on these tasks, existing neural approaches normally require a large amount of labeled training data, which is often unavailable for specific domains such as the Chinese medical domain due to privacy and legal issues. To address this problem, we have developed a Chinese medical corpus named ChiMST which consists of question-answer pairs collected from an online medical healthcare platform and is annotated with word boundary and medical term information. For word boundary, we mainly follow the word segmentation guidelines for the Penn Chinese Treebank (Xia, 2000); for medical terms, we define 9 categories and 18 sub-categories after consulting medical experts. To provide baselines on this corpus, we train existing state-of-the-art models on it and achieve good performance. We believe that the corpus and the baseline systems will be a valuable resource for CWS and NER research on the medical domain.},
  url   = {https://aclanthology.org/2022.lrec-1.607}
}

@InProceedings{singharoy-mercer:2022:LREC,
  author   = {Singha Roy, Sudipta and Mercer, Robert E.},
  title   = {Building a Synthetic Biomedical Research Article Citation Linkage Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5665--5672},
  abstract = {Citations are frequently used in publications to support the presented results and to demonstrate the previous

discoveries while also assisting the reader in following the chronological progression of information through publications. In scientific publications, a citation refers to the referenced document, but it makes no mention of the exact span of text that is being referred to. Connecting the citation to this span of text is called citation linkage. In this paper, to find these citation linkages in biomedical research publications using deep learning, we provide a synthetic silver standard corpus as well as the method to build this corpus. The motivation for building this corpus is to provide a training set for deep learning models that will locate the text spans in a reference article, given a citing statement, based on semantic similarity. This corpus is composed of sentence pairs, where one sentence in each pair is the citing statement and the other one is a candidate cited statement from the referenced paper. The corpus is annotated using an unsupervised sentence embedding method. The effectiveness of this silver standard corpus for training citation linkage models is validated against a human-annotated gold standard corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.608}
}

@InProceedings{kobayashi-EtAl:2022:LREC,
  author   = {Kobayashi, Keita and Koyama, Kohei and Narimatsu, Hiromi and Minami, Yasuhiro},
  title   = {Dataset Construction for Scientific-Document Writing Support by Extracting Related Work Section and Citations from PDF Papers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5673--5682},
  abstract = {To augment datasets used for scientific-document writing support research, we extract texts from ``Related Work'' sections and citation information in PDF-formatted papers published in English. The previous dataset was constructed entirely with Tex-formatted papers, from which it is easy to extract citation information. However, since many publicly available papers in various fields are provided only in PDF format, a dataset constructed using only Tex papers has limited utility. To resolve this problem, we augment the existing dataset by extracting the titles of sections using the visual features of PDF documents and extracting the Related Work section text using the explicit title information. Since text generated from the figures and footnotes appearing in the extraction target areas is considered noise, we remove instances of such text. Moreover, we map the cited paper's information obtained using existing tools to citation marks detected by regular expression rules, resulting in pairs of cited paper information and text of the Related Work section. By evaluating body text extraction and citation mapping in the constructed dataset, the accuracy of the proposed dataset was found to be close to that of the previous dataset. Accordingly, we demonstrated the possibility

of building a significantly augmented dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.609}
}

@InProceedings{debruyne-EtAl:2022:LREC,
  author   = {De Bruyne, Luna and Karimi, Akbar and De Clercq, Orphee and Prati, Andrea and Hoste, Veronique},
  title   = {Aspect-Based Emotion Analysis and Multimodal Coreference: A Case Study of Customer Comments on Adidas Instagram Posts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {574--580},
  abstract = {While aspect-based sentiment analysis of user-generated content has received a lot of attention in the past years, emotion detection at the aspect level has been relatively unexplored. Moreover, given the rise of more visual content on social media platforms, we want to meet the ever-growing share of multimodal content. In this paper, we present a multimodal dataset for Aspect-Based Emotion Analysis (ABEA). Additionally, we take the first steps in investigating the utility of multimodal coreference resolution in an ABEA framework. The presented dataset consists of 4,900 comments on 175 images and is annotated with aspect and emotion categories and the emotional dimensions of valence and arousal. Our preliminary experiments suggest that ABEA does not benefit from multimodal coreference resolution, and that aspect and emotion classification only requires textual information. However, when more specific information about the aspects is desired, image recognition could be essential.},
  url   = {https://aclanthology.org/2022.lrec-1.61}
}

@InProceedings{martynov-EtAl:2022:LREC,
  author   = {Martynov, Nikita and Krotova, Irina and Logacheva, Varvara and Panchenko, Alexander and Kozlova, Olga and Semenov, Nikita},
  title   = {RuPAWS: A Russian Adversarial Dataset for Paraphrase Identification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5683--5691},
  abstract = {Paraphrase identification task can be easily challenged by changing word order, e.g. as in “Can a good person become bad?”. While for English this problem was tackled by the PAWS dataset (Zhang et al., 2019), datasets for Russian paraphrase detection lack non-paraphrase examples with high lexical overlap. We

present RuPAWS, the first adversarial dataset for Russian paraphrase identification. Our dataset consists of examples from PAWS translated to the Russian language and manually annotated by native speakers. We compare it to the largest available dataset for Russian ParaPhraser and show that the best available paraphrase identifiers for the Russian language fail on the RuPAWS dataset. At the same time, the state-of-the-art paraphrasing model RuBERT trained on both RuPAWS and ParaPhraser obtains high performance on the RuPAWS dataset while maintaining its accuracy on the ParaPhraser benchmark. We also show that RuPAWS can measure the sensitivity of models to word order and syntax structure since simple baselines fail even when given RuPAWS training samples.},
  url   = {https://aclanthology.org/2022.lrec-1.610}
}

@InProceedings{rodriguesgomide-carapinha-plag:2022:LREC,
  author   = {Rodrigues Gomide, Andressa and Carapinha, Conceição and Plag, Cornelia},
  title   = {Atril: an XML Visualization System for Corpus Texts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5692--5695},
  abstract = {This paper presents Atril, an XML visualization system for corpus texts, developed for, but not restricted to, the project Corpus de Audiências (CorAuDis), a corpus composed of transcripts of sessions of criminal proceedings recorded at the Coimbra Court. The main aim of the tool is to provide researchers with a web-based environment that allows for an easily customizable visualization of corpus texts with heavy structural annotation. Existing corpus analysis tools such as SketchEngine, TEITOK and CQPweb offer some kind of visualization mechanisms, but, to our knowledge, none meets our project’s main needs. Our requirements are a system that is open-source; that can be easily connected to CQPweb and TEITOK, that provides a full text-view with switchable visualization templates, that allows for the visualization of overlapping utterances. To meet those requirements, we created Atril, a module with a corpus XML file viewer, a visualization management system, and a word alignment tool.},
  url   = {https://aclanthology.org/2022.lrec-1.611}
}

@InProceedings{arora-venkateswaran-schneider:2022:LREC,
  author   = {Arora, Aryaman and Venkateswaran, Nitin and Schneider, Nathan},
  title   = {MASALA: Modelling and Analysing the Semantics of Adpositions in Linguistic Annotation of Hindi},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5696--5704},
  abstract = {We present a completed, publicly available corpus of annotated semantic relations of adpositions and case markers in Hindi. We used the multilingual SNACS annotation scheme, which has been applied to a variety of typologically diverse languages. Building on past work examining linguistic problems in SNACS annotation, we use language models to attempt automatic labelling of SNACS supersenses in Hindi and achieve results competitive with past work on English. We look towards upstream applications in semantic role labelling and extension to related languages such as Gujarati.},
  url   = {https://aclanthology.org/2022.lrec-1.612}
}

@InProceedings{arora:2022:LREC,
  author   = {Arora, Aryaman},
  title   = {Universal Dependencies for Punjabi},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5705--5711},
  abstract = {We introduce the first Universal Dependencies treebank for Punjabi (written in the Gurmukhi script) and discuss corpus design and linguistic phenomena encountered in annotation. The treebank covers a variety of genres and has been annotated for POS tags, dependency relations, and graph-based Enhanced Dependencies. We aim to expand the diversity of coverage of Indo-Aryan languages in UD.},
  url   = {https://aclanthology.org/2022.lrec-1.613}
}

@InProceedings{urlana-EtAl:2022:LREC,
  author   = {Urlana, Ashok and Surange, Nirmal and Baswani, Pavan and Ravva, Priyanka and Shrivastava, Manish},
  title   = {TeSum: Human-Generated Abstractive Summarization Corpus for Telugu},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5712--5722},
  abstract = {Expert human annotation for summarization is definitely an expensive task, and can not be done on huge scales. But with this work, we show that even with a crowd sourced summary generation approach, quality can be controlled by aggressive expert informed filtering and sampling-based human evaluation. We propose a pipeline that crowd-sources summarization data and then aggressively

filters the content via: automatic and partial expert evaluation. Using this pipeline we create a high-quality Telugu Abstractive Summarization dataset (TeSum) which we validate with sampling-based human evaluation. We also provide baseline numbers for various models commonly used for summarization. A number of recently released datasets for summarization, scraped the web-content relying on the assumption that summary is made available with the article by the publishers. While this assumption holds for multiple resources (or news-sites) in English, it should not be generalised across languages without thorough analysis and verification. Our analysis clearly shows that this assumption does not hold true for most Indian language news resources. We show that our proposed filtration pipeline can even be applied to these large-scale scraped datasets to extract better quality article-summary pairs.},
  url   = {https://aclanthology.org/2022.lrec-1.614}
}

@InProceedings{lee-EtAl:2022:LREC3,
  author   = {Lee, John and Fong, Haley and Wong, Lai Shuen Judy and Mak, Chun Chung and Yip, Chi Hin and Ng, Ching Wah Larry},
  title   = {A Corpus of Simulated Counselling Sessions with Dialog Act Annotation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5723--5730},
  abstract = {We present a corpus of simulated counselling sessions consisting of speech- and text-based dialogs in Cantonese. Consisting of 152K Chinese characters, the corpus labels the dialog act of both client and counsellor utterances, segments each dialog into stages, and identifies the forward and backward links in the dialog. We analyze the distribution of client and counsellor communicative intentions in the various stages, and discuss significant patterns of the dialog flow.},
  url   = {https://aclanthology.org/2022.lrec-1.615}
}

@InProceedings{mehri-EtAl:2022:LREC,
  author   = {Mehri, Shikib and Feng, Yulan and Gordon, Carla and Alavi, Seyed Hossein and Traum, David and Eskenazi, Maxine},
  title   = {Interactive Evaluation of Dialog Track at DSTC9},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5731--5738},
  abstract = {The ultimate goal of dialog research is to develop

systems that can be effectively used in interactive settings by real users. To this end, we introduced the Interactive Evaluation of Dialog Track at the 9th Dialog System Technology Challenge. This track consisted of two sub-tasks. The first sub-task involved building knowledge-grounded response generation models. The second sub-task aimed to extend dialog models beyond static datasets by assessing them in an interactive setting with real users. Our track challenges participants to develop strong response generation models and explore strategies that extend them to back-and-forth interactions with real users. The progression from static corpora to interactive evaluation introduces unique challenges and facilitates a more thorough assessment of open-domain dialog systems. This paper provides an overview of the track, including the methodology and results. Furthermore, it provides insights into how to best evaluate open-domain dialog models.},
  url   = {https://aclanthology.org/2022.lrec-1.616}
}

@InProceedings{torresfonsesca-kennington:2022:LREC,
  author   = {Torres-Fonsesca, Josue and Kennington, Casey},
  title   = {HADREB: Human Appraisals and (English) Descriptions of Robot Emotional Behaviors},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5739--5748},
  abstract = {Humans sometimes anthropomorphize everyday objects, but especially robots that have human-like qualities and that are often able to interact with and respond to humans in ways that other objects cannot. Humans especially attribute emotion to robot behaviors, partly because humans often use and interpret emotions when interacting with other humans, and they apply that capability when interacting with robots. Moreover, emotions are a fundamental part of the human language system and emotions are used as scaffolding for language learning, making them an integral part of language learning and meaning. However, there are very few datasets that explore how humans perceive the emotional states of robots and how emotional behaviors relate to human language. To address this gap we have collected HADREB, a dataset of human appraisals and English descriptions of robot emotional behaviors collected from over 30 participants. These descriptions and human emotion appraisals are collected using the Mistyrobotics Misty II and the Digital Dream Labs Cozmo (formerly Anki) robots. The dataset contains English descriptions and emotion appraisals of more than 500 descriptions and graded valence labels of 8 emotion pairs for each behavior and each robot. In this paper we describe the process of collecting and cleaning the data, give a general analysis of the data, and evaluate the usefulness of the dataset in two experiments, one using a language model to map descriptions to emotions, the other maps robot behaviors to emotions.},
  url   = {https://aclanthology.org/2022.lrec-1.617}

}

@InProceedings{mitsuda-EtAl:2022:LREC,
  author   = {Mitsuda, Koh and Higashinaka, Ryuichiro and Oga, Yuhei and Yoshida, Sen},
  title   = {Dialogue Collection for Recording the Process of Building Common Ground in a Collaborative Task},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5749--5758},
  abstract = {To develop a dialogue system that can build common ground with users, the process of building common ground through dialogue needs to be clarified. However, the studies on the process of building common ground have not been well conducted; much work has focused on finding the relationship between a dialogue in which users perform a collaborative task and its task performance represented by the final result of the task. In this study, to clarify the process of building common ground, we propose a data collection method for automatically recording the process of building common ground through a dialogue by using the intermediate result of a task. We collected 984 dialogues, and as a result of investigating the process of building common ground, we found that the process can be classified into several typical patterns and that conveying each worker's understanding through affirmation of a counterpart's utterances especially contributes to building common ground. In addition, toward dialogue systems that can build common ground, we conducted an automatic estimation of the degree of built common ground and found that its degree can be estimated quite accurately.},
  url   = {https://aclanthology.org/2022.lrec-1.618}
}

@InProceedings{inaba-EtAl:2022:LREC,
  author   = {Inaba, Michimasa and Chiba, Yuya and Higashinaka, Ryuichiro and Komatani, Kazunori and Miyao, Yusuke and Nagai, Takayuki},
  title   = {Collection and Analysis of Travel Agency Task Dialogues with Age-Diverse Speakers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5759--5767},
  abstract = {When individuals communicate with each other, they use different vocabulary, speaking speed, facial expressions, and body language depending on the people they talk to. This paper focuses on the speaker's age as a factor that affects the change in communication. We collected a multimodal dialogue corpus with a wide

range of speaker ages. As a dialogue task, we focus on travel, which interests people of all ages, and we set up a task based on a tourism consultation between an operator and a customer at a travel agency. This paper provides details of the dialogue task, the collection procedure and annotations, and the analysis on the characteristics of the dialogues and facial expressions focusing on the age of the speakers. Results of the analysis suggest that the adult speakers have more independent opinions, the older speakers more frequently express their opinions frequently compared with other age groups, and the operators expressed a smile more frequently to the minor speakers.},
  url   = {https://aclanthology.org/2022.lrec-1.619}
}

@InProceedings{roccabruna-azzolin-riccardi:2022:LREC,
  author   = {Roccabruna, Gabriel and Azzolin, Steve and Riccardi, Giuseppe},
  title   = {Multi-source Multi-domain Sentiment Analysis with BERT-based Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {581--589},
  abstract = {Sentiment analysis is one of the most widely studied tasks in natural language processing. While BERT-based models have achieved state-of-the-art results in this task, little attention has been given to its performance variability across class labels, multi-source and multi-domain corpora. In this paper, we present an improved state-of-the-art and comparatively evaluate BERT-based models for sentiment analysis on Italian corpora. The proposed model is evaluated over eight sentiment analysis corpora from different domains (social media, finance, e-commerce, health, travel) and sources (Twitter, YouTube, Facebook, Amazon, Tripadvisor, Opera and Personal Healthcare Agent) on the prediction of positive, negative and neutral classes. Our findings suggest that BERT-based models are confident in predicting positive and negative examples but not as much with neutral examples. We release the sentiment analysis model as well as a newly financial domain sentiment corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.62}
}

@InProceedings{karkada-EtAl:2022:LREC,
  author   = {Karkada, Deepthi and Manuvinakurike, Ramesh and Paetzel-Prüsmann, Maike and Georgila, Kallirroi},
  title   = {Strategy-level Entrainment of Dialogue System Users in a Creative Visual Reference Resolution Task},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

publisher = {European Language Resources Association},
pages = {5768--5777},
abstract = {In this work, we study entrainment of users playing a creative reference resolution game with an autonomous dialogue system. The language understanding module in our dialogue system leverages annotated human-wizard conversational data, openly available knowledge graphs, and crowd-augmented data. Unlike previous entrainment work, our dialogue system does not attempt to make the human conversation partner adopt lexical items in their dialogue, but rather to adapt their descriptive strategy to one that is simpler to parse for our natural language understanding unit. By deploying this dialogue system through a crowd-sourced study, we show that users indeed entrain on a "strategy-level" without the change of strategy impinging on their creativity. Our work thus presents a promising future research direction for developing dialogue management systems that can strategically influence people’s descriptive strategy to ease the system’s language understanding in creative tasks.},
url = {https://aclanthology.org/2022.lrec-1.620}
}

@InProceedings{zheng-EtAl:2022:LREC,
author = {Zheng, Yinhe and Chen, Guanyi and Liu, Xin and Sun, Jian},
title = {MMChat: Multi-Modal Chat Dataset on Social Media},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {5778--5786},
abstract = {Incorporating multi-modal contexts in conversation is an important step for developing more engaging dialogue systems. In this work, we explore this direction by introducing MMChat: a large scale Chinese multi-modal dialogue corpus (32.4M raw dialogues and 120.84K filtered dialogues). Unlike previous corpora that are crowd-sourced or collected from fictitious movies, MMChat contains image-grounded dialogues collected from real conversations on social media, in which the sparsity issue is observed. Specifically, image-initiated dialogues in common communications may deviate to some non-image-grounded topics as the conversation proceeds. To better investigate this issue, we manually annotate 100K dialogues from MMChat and further filter the corpus accordingly, which yields MMChat-hf. We develop a benchmark model to address the sparsity issue in dialogue generation tasks by adapting the attention routing mechanism on image features. Experiments demonstrate the usefulness of incorporating image features and the effectiveness in handling the sparsity of image features.},
url = {https://aclanthology.org/2022.lrec-1.621}
}

@InProceedings{jia-EtAl:2022:LREC1,
author = {jia, meihuizi and Liu, Ruixue and Wang, Peiying

and Song, Yang and Xi, Zexi and Li, Haobin and Shen, Xin and Chen, Meng and Pang, Jinhui and He, Xiaodong},
  title = {E-ConvRec: A Large-Scale Conversational Recommendation Dataset for E-Commerce Customer Service},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {5787--5796},
  abstract = {There has been a growing interest in developing conversational recommendation system (CRS), which provides valuable recommendations to users through conversations. Compared to the traditional recommendation, it advocates wealthier interactions and provides possibilities to obtain users' exact preferences explicitly. Nevertheless, the corresponding research on this topic is limited due to the lack of broad-coverage dialogue corpus, especially real-world dialogue corpus. To handle this issue and facilitate our exploration, we construct E-ConvRec, an authentic Chinese dialogue dataset consisting of over 25k dialogues and 770k utterances, which contains user profile, product knowledge base (KB), and multiple sequential real conversations between users and recommenders. Next, we explore conversational recommendation in a real scene from multiple facets based on the dataset. Therefore, we particularly design three tasks: user preference recognition, dialogue management, and personalized recommendation. In the light of the three tasks, we establish baseline results on E-ConvRec to facilitate future studies.},
  url = {https://aclanthology.org/2022.lrec-1.622}
}

@InProceedings{monsur-EtAl:2022:LREC,
  author = {Monsur, Syed Mostofa and Chowdhury, Sakib and Fatemi, Md Shahrar and Ahmed, Shafayat},
  title = {SHONGLAP: A Large Bengali Open-Domain Dialogue Corpus},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {5797--5804},
  abstract = {We introduce SHONGLAP, a large annotated open-domain dialogue corpus in Bengali language. Due to unavailability of high-quality dialogue datasets for low-resource languages like Bengali, existing neural open-domain dialogue systems suffer from data scarcity. We propose a framework to prepare large-scale open-domain dialogue datasets from publicly available multi-party discussion podcasts, talk-shows and label them based on weak-supervision techniques which is particularly suitable for low-resource settings. Using this framework, we prepared our corpus, the first reported Bengali open-domain dialogue corpus (7.7k+ fully annotated dialogues

in total) which can serve as a strong baseline for future works. Experimental results show that our corpus improves performance of large language models (BanglaBERT) in case of downstream classification tasks during fine-tuning.},
  url   = {https://aclanthology.org/2022.lrec-1.623}
}

@InProceedings{onishi-EtAl:2022:LREC,
  author   = {Onishi, Toshiki and Ogushi, Asahi and Tahara, Yohei and Ishii, Ryo and Fukayama, Atsushi and Nakamura, Takao and Miyata, Akihiro},
  title   = {A Comparison of Praising Skills in Face-to-Face and Remote Dialogues},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5805--5812},
  abstract = {Praising behavior is considered to an important method of communication in daily life and social activities. An engineering analysis of praising behavior is therefore valuable. However, a dialogue corpus for this analysis has not yet been developed. Therefore, we develop corpuses for face-to-face and remote two-party dialogues with ratings of praising skills. The corpuses enable us to clarify how to use verbal and nonverbal behaviors for successfully praise. In this paper, we analyze the differences between the face-to-face and remote corpuses, in particular the expressions in adjudged praising scenes in both corpuses, and also evaluated praising skills. We also compare differences in head motion, gaze behavior, facial expression in high-rated praising scenes in both corpuses. The results showed that the distribution of praising scores was similar in face-to-face and remote dialogues, although the ratio of the number of praising scenes to the number of utterances was different. In addition, we confirmed differences in praising behavior in face-to-face and remote dialogues.},
  url   = {https://aclanthology.org/2022.lrec-1.624}
}

@InProceedings{tur-traum:2022:LREC,
  author   = {Tur, Ada and Traum, David},
  title   = {Comparing Approaches to Language Understanding for Human-Robot Dialogue: An Error Taxonomy and Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5813--5820},
  abstract = {In this paper, we compare two different approaches to language understanding for a human-robot interaction domain in which

a human commander gives navigation instructions to a robot. We contrast a relevance-based classifier with a GPT-2 model, using about 2000 input-output examples as training data. With this level of training data, the relevance-based model outperforms the GPT-2 based model 79\% to \68\%. We also present a taxonomy of types of errors made by each model, indicating that they have somewhat different strengths and weaknesses, so we also examine the potential for a combined model.},
  url   = {https://aclanthology.org/2022.lrec-1.625}
}

@InProceedings{sun-cao-yang:2022:LREC,
  author   = {Sun, Hanfei and Cao, Ziyuan and Yang, Diyi},
  title   = {SPORTSINTERVIEW: A Large-Scale Sports Interview Benchmark for Entity-centric Dialogues},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5821--5828},
  abstract = {We propose a novel knowledge grounded dialogue (interview) dataset SPORTSINTERVIEW set in the domain of sports interview. Our dataset contains two types of external knowledge sources as knowledge grounding, and is rich in content, containing about 150K interview sessions and 34K distinct interviewees. Compared to existing knowledge grounded dialogue datasets, our interview dataset is larger in size, comprises natural dialogues revolving around real-world sports matches, and have more than one dimension of external knowledge linking. We performed several experiments on SPORTSINTERVIEW and found that models such as BART fine-tuned on our dataset are able to learn lots of relevant domain knowledge and generate meaningful sentences (questions or responses). However, their performance is still far from humans (by comparing to gold sentences in the dataset) and hence encourages future research utilizing SPORTSINTERVIEW.},
  url   = {https://aclanthology.org/2022.lrec-1.626}
}

@InProceedings{singh-EtAl:2022:LREC3,
  author   = {Singh, Gopendra Vikram and Priya, Priyanshu and Firdaus, Mauajama and Ekbal, Asif and Bhattacharyya, Pushpak},
  title   = {EmoInHindi: A Multi-label Emotion and Intensity Annotated Dataset in Hindi for Emotion Recognition in Dialogues},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5829--5837},
  abstract = {The long-standing goal of Artificial Intelligence (AI) has been to create human-like conversational systems. Such

systems should have the ability to develop an emotional connection with the users, consequently, emotion recognition in dialogues has gained popularity. Emotion detection in dialogues is a challenging task because humans usually convey multiple emotions with varying degrees of intensities in a single utterance. Moreover, emotion in an utterance of a dialogue may be dependent on previous utterances making the task more complex. Recently, emotion recognition in low-resource languages like Hindi has been in great demand. However, most of the existing datasets for multi-label emotion and intensity detection in conversations are in English. To this end, we propose a large conversational dataset in Hindi named EmoInHindi for multi-label emotion and intensity recognition in conversations containing 1,814 dialogues with a total of 44,247 utterances. We prepare our dataset in a Wizard-of-Oz manner for mental health and legal counselling of crime victims. Each utterance of dialogue is annotated with one or more emotion categories from 16 emotion labels including neutral and their corresponding intensity. We further propose strong contextual baselines that can detect the emotion(s) and corresponding emotional intensity of an utterance given the conversational context.},
  url = {https://aclanthology.org/2022.lrec-1.627}
}

@InProceedings{vishnubhotla-hammond-hirst:2022:LREC,
  author = {Vishnubhotla, Krishnapriya and Hammond, Adam and Hirst, Graeme},
  title = {The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {5838--5848},
  abstract = {We present the Project Dialogism Novel Corpus, or PDNC, an annotated dataset of quotations for English literary texts. PDNC contains annotations for 35,978 quotations across 22 full-length novels, and is by an order of magnitude the largest corpus of its kind. Each quotation is annotated for the speaker, addressees, type of quotation, referring expression, and character mentions within the quotation text. The annotated attributes allow for a comprehensive evaluation of models of quotation attribution and coreference for literary texts.},
  url = {https://aclanthology.org/2022.lrec-1.628}
}

@InProceedings{rehbein-ruppenhofer:2022:LREC,
  author = {Rehbein, Ines and Ruppenhofer, Josef},
  title = {Who’s in, who’s out? Predicting the Inclusiveness or Exclusiveness of Personal Pronouns in Parliamentary Debates},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5849--5858},
  abstract = {This paper presents a compositional annotation scheme to capture the clusivity properties of personal pronouns in context, that is their ability to construct and manage in-groups and out-groups by including/excluding the audience and/or non-speech act participants in reference to groups that also include the speaker. We apply and test our schema on pronoun instances in speeches taken from the German parliament. The speeches cover a time period from 2017-2021 and comprise manual annotations for 3,126 sentences. We achieve high inter-annotator agreement for our new schema, with a Cohen's κ in the range of 89.7-93.2 and a percentage agreement of > 96\%. Our exploratory analysis of in/exclusive pronoun use in the parliamentary setting provides some face validity for our new schema. Finally, we present baseline experiments for automatically predicting clusivity in political debates, with promising results for many referential constellations, yielding an overall 84.9\% micro F1 for all pronouns.},
  url   = {https://aclanthology.org/2022.lrec-1.629}
}

@InProceedings{muhammad-EtAl:2022:LREC,
  author   = {Muhammad, Shamsuddeen Hassan and Adelani, David and Aremu, Anuoluwapo and Abdulmumin, Idris},
  title   = {NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {590--602},
  abstract = {Sentiment analysis is one of the most widely studied applications in NLP, but most work focuses on languages with large amounts of data. We introduce the first large-scale human-annotated Twitter sentiment dataset for the four most widely spoken languages in Nigeria—Hausa, Igbo, Nigerian-Pidgin, and Yorùbá—consisting of around 30,000 annotated tweets per language, including a significant fraction of code-mixed tweets. We propose text collection, filtering, processing and labeling methods that enable us to create datasets for these low-resource languages. We evaluate a range of pre-trained models and transfer strategies on the dataset. We find that language-specific models and language-adaptive fine-tuning generally perform best. We release the datasets, trained models, sentiment lexicons, and code to incentivize research on sentiment analysis in under-represented languages.},
  url   = {https://aclanthology.org/2022.lrec-1.63}
}

@InProceedings{booth-shoemaker-gaizauskas:2022:LREC,
  author   = {Booth, Callum and Shoemaker, Robert and

Gaizauskas, Robert},
  title   = {A Language Modelling Approach to Quality Assessment of OCR'ed Historical Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5859--5864},
  abstract = {We hypothesise and evaluate a language model-based approach for scoring the quality of OCR transcriptions in the British Library Newspapers (BLN) corpus parts 1 and 2, to identify the best quality OCR for use in further natural language processing tasks, with a wider view to link individual newspaper reports of crime in nineteenth-century London to the Digital Panopticon---a structured repository of criminal lives. We mitigate the absence of gold standard transcriptions of the BLN corpus by utilising a corpus of genre-adjacent texts that capture the common and legal parlance of nineteenth-century London---the Proceedings of the Old Bailey Online---with a view to rank the BLN transcriptions by their OCR quality.},
  url   = {https://aclanthology.org/2022.lrec-1.630}
}

@InProceedings{morante-EtAl:2022:LREC,
  author   = {Morante, Roser and Smith, Eleanor L. T. and Wilhelmus, Lianne and Lassche, Alie and Kuijpers, Erika},
  title   = {Identifying Copied Fragments in a 18th Century Dutch Chronicle},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5865--5878},
  abstract = {We apply computational stylometric techniques to an 18th century Dutch chronicle to determine which fragments of the manuscript represent the author's own original work and which show signs of external source use through either direct copying or paraphrasing. Through stylometric methods the majority of text fragments in the chronicle can be correctly labelled as either the author's own words, direct copies from sources or paraphrasing. Our results show that clustering text fragments based on stylometric measures is an effective methodology for authorship verification of this document; however, this approach is less effective when personal writing style is masked by author independent styles or when applied to paraphrased text.},
  url   = {https://aclanthology.org/2022.lrec-1.631}
}

@InProceedings{liagkou-pavlopoulos-machotka:2022:LREC,
  author   = {Liagkou, Konstantina and Pavlopoulos, John and

Machotka, Ewa},
  title   = {A Study of Distant Viewing of ukiyo-e prints},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5879--5888},
  abstract = {This paper contributes to studying relationships between Japanese topography and places featured in early modern landscape prints, so-called ukiyo-e or ‘pictures of the floating world’. The printed inscriptions on these images feature diverse place-names, both man-made and natural formations. However, due to the corpus's richness and diversity, the precise nature of artistic mediation of the depicted places remains little understood. In this paper, we explored a new analytical approach based on the macroanalysis of images facilitated by Natural Language Processing technologies. This paper presents a small dataset with inscriptions on prints that have been annotated by an art historian for included place-name entities. Our dataset is released for public use. By fine-tuning and applying a Japanese BERT-based Name Entity Recogniser, we provide a use-case of a macroanalysis of a visual dataset that is hosted by the digital database of the Art Research Center at the Ritsumeikan University, Kyoto. Our work studies the relationship between topography and its visual renderings in early modern Japanese ukiyo-e landscape prints, demonstrating how an art historian’s work can be improved with Natural Language Processing toward distant viewing of visual datasets. We release our dataset and code for public use: https://github.com/connalia/ukiyo-e_meisho_nlp},
  url   = {https://aclanthology.org/2022.lrec-1.632}
}

@InProceedings{wang-riddell:2022:LREC,
  author   = {Wang, Haining and Riddell, Allen},
  title   = {CCTAA: A Reproducible Corpus for Chinese Authorship Attribution Research},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5889--5893},
  abstract = {Authorship attribution infers the likely author of an unsigned, single-authored document from a pool of candidates. Despite recent advances, a lack of standard, reproducible testbeds for Chinese language documents impedes progress. In this paper, we present the Chinese Cross-Topic Authorship Attribution (CCTAA) corpus. It is the first standard testbed for authorship attribution on contemporary Chinese prose. The cross-topic design and relatively inflexible genre of newswire contribute to an appropriate level of difficulty. It supports reproducible research by using pre-defined

data splits. We show that a sequence classifier based on pre-trained Chinese RoBERTa embedding and a support vector machine classifier using function character n-gram frequency features perform below expectations on this task. The code for generating the corpus and reproducing the baselines is freely available at https://codeberg.org/haining/cctaa.},
  url   = {https://aclanthology.org/2022.lrec-1.633}
}

@InProceedings{yousef-EtAl:2022:LREC,
  author   = {Yousef, Tariq and Palladino, Chiara and Shamsian, Farnoosh and d'Orange Ferreira, Anise and Ferreira dos Reis, Michel},
  title   = {An automatic model and Gold Standard for translation alignment of Ancient Greek},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5894--5905},
  abstract = {This paper illustrates a workflow for developing and evaluating automatic translation alignment models for Ancient Greek. We designed an annotation Style Guide and a gold standard for the alignment of Ancient Greek-English and Ancient Greek-Portuguese, measured inter-annotator agreement and used the resulting dataset to evaluate the performance of various translation alignment models. We proposed a fine-tuning strategy that employs unsupervised training with mono- and bilingual texts and supervised training using manually aligned sentences. The results indicate that the fine-tuned model based on XLM-Roberta is superior in performance, and it achieved good results on language pairs that were not part of the training data.},
  url   = {https://aclanthology.org/2022.lrec-1.634}
}

@InProceedings{vargas-EtAl:2022:LREC1,
  author   = {Vargas, Francielle and D`Alessandro, Jonas and Rabinovich, Zohar and Benevenuto, Fabrício and Pardo, Thiago},
  title   = {Rhetorical Structure Approach for Online Deception Detection: A Survey},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5906--5915},
  abstract = {Most information is passed on in the form of language. Therefore, research on how people use language to inform and misinform, and how this knowledge may be automatically extracted from large amounts of text is surely relevant. This survey provides first-hand experiences and a comprehensive review of rhetorical-

level structure analysis for online deception detection. We systematically analyze how discourse structure, aligned or not with other approaches, is applied to automatic fake news and fake reviews detection on the web and social media. Moreover, we categorize discourse-tagged corpora along with results, hence offering a summary and accessible introductions to new researchers.},
  url   = {https://aclanthology.org/2022.lrec-1.635}
}

@InProceedings{naito-EtAl:2022:LREC,
  author   = {Naito, Shoichi and Sawada, Shintaro and Nakagawa, Chihiro and Inoue, Naoya and Yamaguchi, Kenshi and Shimizu, Iori and Mim, Farjana Sultana and Singh, Keshav and Inui, Kentaro},
  title   = {TYPIC: A Corpus of Template-Based Diagnostic Comments on Argumentation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5916--5928},
  abstract = {Providing feedback on the argumentation of the learner is essential for developing critical thinking skills, however, it requires a lot of time and effort. To mitigate the overload on teachers, we aim to automate a process of providing feedback, especially giving diagnostic comments which point out the weaknesses inherent in the argumentation. It is recommended to give specific diagnostic comments so that learners can recognize the diagnosis without misinterpretation. However, it is not obvious how the task of providing specific diagnostic comments should be formulated. We present a formulation of the task as template selection and slot filling to make an automatic evaluation easier and the behavior of the model more tractable. The key to the formulation is the possibility of creating a template set that is sufficient for practical use. In this paper, we define three criteria that a template set should satisfy: expressiveness, informativeness, and uniqueness, and verify the feasibility of creating a template set that satisfies these criteria as a first trial. We will show that it is feasible through an annotation study that converts diagnostic comments given in a text to a template format. The corpus used in the annotation study is publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.636}
}

@InProceedings{mendonca-EtAl:2022:LREC,
  author   = {Mendonca, John and Correia, Rui and Lourenço, Mariana and Freitas, João and Trancoso, Isabel},
  title   = {Towards Speaker Verification for Crowdsourced Speech Collections},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5929--5937},
  abstract = {Crowdsourcing the collection of speech provides a scalable setting to access a customisable demographic according to each dataset's needs. The correctness of speaker metadata is especially relevant for speaker-centred collections - ones that require the collection of a fixed amount of data per speaker. This paper identifies two different types of misalignment present in these collections: Multiple Accounts misalignment (different contributors map to the same speaker), and Multiple Speakers misalignment (multiple speakers map to the same contributor). Based on state-of-the-art approaches to Speaker Verification, this paper proposes an unsupervised method for measuring speaker metadata plausibility of a collection, i.e., evaluating the match (or lack thereof) between contributors and speakers. The solution presented is composed of an embedding extractor and a clustering module. Results indicate high precision in automatically classifying contributor alignment (>0.94).},
  url   = {https://aclanthology.org/2022.lrec-1.637}
}

@InProceedings{xiao-EtAl:2022:LREC,
  author   = {Xiao, Liming and Li, Bin and Xu, Zhixing and Huo, Kairui and Feng, Minxuan and Zhou, Junsheng and Qu, Weiguang},
  title   = {Align-smatch: A Novel Evaluation Method for Chinese Abstract Meaning Representation Parsing based on Alignment of Concept and Relation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5938--5945},
  abstract = {Abstract Meaning Representation is a sentence-level meaning representation, which abstracts the meaning of sentences into a rooted acyclic directed graph. With the continuous expansion of Chinese AMR corpus, more and more scholars have developed parsing systems to automatically parse sentences into Chinese AMR. However, the current parsers can't deal with concept alignment and relation alignment, let alone the evaluation methods for AMR parsing. Therefore, to make up for the vacancy of Chinese AMR parsing evaluation methods, based on AMR evaluation metric smatch, we have improved the algorithm of generating triples so that to make it compatible with concept alignment and relation alignment. Finally, we obtain a new integrity metric align-smatch for paring evaluation. A comparative research then was conducted on 20 manually annotated AMR and gold AMR, with the result that align-smatch works well in alignments and more robust in evaluating arcs. We also put forward some fine-grained metric for evaluating concept alignment, relation

alignment and implicit concepts, in order to further measure parsers' performance in subtasks.},
  url   = {https://aclanthology.org/2022.lrec-1.638}
}

@InProceedings{thorleiksdttir-EtAl:2022:LREC,
  author   = {Thorleiksdóttir, Thórhildur and Renggli, Cedric and Hollenstein, Nora and Zhang, Ce},
  title   = {Dynamic Human Evaluation for Relative Model Comparisons},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5946--5955},
  abstract = {Collecting human judgements is currently the most reliable evaluation method for natural language generation systems. Automatic metrics have reported flaws when applied to measure quality aspects of generated text and have been shown to correlate poorly with human judgements. However, human evaluation is time and cost-intensive, and we lack consensus on designing and conducting human evaluation experiments. Thus there is a need for streamlined approaches for efficient collection of human judgements when evaluating natural language generation systems. Therefore, we present a dynamic approach to measure the required number of human annotations when evaluating generated outputs in relative comparison settings. We propose an agent-based framework of human evaluation to assess multiple labelling strategies and methods to decide the better model in a simulation and a crowdsourcing case study. The main results indicate that a decision about the superior model can be made with high probability across different labelling strategies, where assigning a single random worker per task requires the least overall labelling effort and thus the least cost.},
  url   = {https://aclanthology.org/2022.lrec-1.639}
}

@InProceedings{etienne-battistelli-lecorv:2022:LREC,
  author   = {Etienne, Aline and Battistelli, Delphine and Lecorvé, Gwénolé},
  title   = {A (Psycho-)Linguistically Motivated Scheme for Annotating and Exploring Emotions in a Genre-Diverse Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {603--612},
  abstract = {This paper presents a scheme for emotion annotation and its manual application on a genre-diverse corpus of texts written in French. The methodology introduced here emphasizes the necessity of clarifying the main concepts implied by the analysis of

emotions as they are expressed in texts, before conducting a manual annotation campaign. After explaining whatentails a deeply linguistic perspective on emotion expression modeling, we present a few NLP works that share some common points with this perspective and meticulously compare our approach with them. We then highlight some interesting quantitative results observed on our annotated corpus. The most notable interactions are on the one hand between emotion expression modes and genres of texts, and on the other hand between emotion expression modes and emotional categories. These observation corroborate and clarify some of the results already mentioned in other NLP works on emotion annotation.},
  url   = {https://aclanthology.org/2022.lrec-1.64}
}

@InProceedings{bestgen:2022:LREC,
  author   = {Bestgen, Yves},
  title   = {Please, Don't Forget the Difference and the Confidence Interval when Seeking for the State-of-the-Art Status},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5956--5962},
  abstract = {This paper argues for the widest possible use of bootstrap confidence intervals for comparing NLP system performances instead of the state-of-the-art status (SOTA) and statistical significance testing. Their main benefits are to draw attention to the difference in performance between two systems and to help assessing the degree of superiority of one system over another. Two cases studies, one comparing several systems and the other based on a K-fold cross-validation procedure, illustrate these benefits.},
  url   = {https://aclanthology.org/2022.lrec-1.640}
}

@InProceedings{zhao-zhang-song:2022:LREC,
  author   = {Zhao, Xinran and Zhang, Hongming and Song, Yangqiu},
  title   = {PCR4ALL: A Comprehensive Evaluation Benchmark for Pronoun Coreference Resolution in English},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5963--5973},
  abstract = {Pronoun Coreference Resolution (PCR) is the task of resolving pronominal expressions to all mentions they refer to. The correct resolution of pronouns typically involves the complex inference over both linguistic knowledge and general world knowledge. Recently, with the help of pre-trained language representation models, the community has made significant progress

on various PCR tasks. However, as most existing works focus on developing PCR models for specific datasets and measuring the accuracy or F1 alone, it is still unclear whether current PCR systems are reliable in real applications. Motivated by this, we propose PCR4ALL, a new benchmark and a toolbox that evaluates and analyzes the performance of PCR systems from different perspectives (i.e., knowledge source, domain, data size, frequency, relevance, and polarity). Experiments demonstrate notable performance differences when the models are examined from different angles. We hope that PCR4ALL can motivate the community to pay more attention to solving the overall PCR problem and understand the performance comprehensively. All data and codes are available at: https://github.com/HKUST-KnowComp/PCR4ALL.},
  url   = {https://aclanthology.org/2022.lrec-1.641}
}

@InProceedings{lepekhin-sharoff:2022:LREC,
  author   = {Lepekhin, Mikhail and Sharoff, Serge},
  title   = {Estimating Confidence of Predictions of Individual Classifiers and TheirEnsembles for the Genre Classification Task},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5974--5982},
  abstract = {Genre identification is a kind of non-topic text classification. The main difference between this task and topic classification is that genre, unlike topic, usually cannot be expressed just by some keywords and is defined as a functional space. Neural models based on pre-trained transformers, such as BERT or XLM-RoBERTa, demonstrate SOTA results in many NLP tasks, including non-topical classification. However, in many cases, their downstream application to very large corpora, such as those extracted from social media, can lead to unreliable results because of dataset shifts, when some raw texts do not match the profile of the training set. To mitigate this problem, we experiment with individual models as well as with their ensembles. To evaluate the robustness of all models we use a prediction confidence metric, which estimates the reliability of a prediction in the absence of a gold standard label. We can evaluate robustness via the confidence gap between the correctly classified texts and the misclassified ones on a labeled test corpus, higher gaps make it easier to identify whether a text is classified correctly. Our results show that for all of the classifiers tested in this study, there is a confidence gap, but for the ensembles, the gap is wider, meaning that ensembles are more robust than their individual models.},
  url   = {https://aclanthology.org/2022.lrec-1.642}
}

@InProceedings{vajjala-balasubramaniam:2022:LREC,
  author   = {Vajjala, Sowmya and Balasubramaniam, Ramya},
  title   = {What do we really know about State of the Art NER?},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5983--5993},
  abstract = {Named Entity Recognition (NER) is a well researched NLP task and is widely used in real world NLP scenarios. NER research typically focuses on the creation of new ways of training NER, with relatively less emphasis on resources and evaluation. Further, state of the art (SOTA) NER models, trained on standard datasets, typically report only a single performance measure (F-score) and we don't really know how well they do for different entity types and genres of text, or how robust are they to new, unseen entities. In this paper, we perform a broad evaluation of NER using a popular dataset, that takes into consideration various text genres and sources constituting the dataset at hand. Additionally, we generate six new adversarial test sets through small perturbations in the original test set, replacing select entities while retaining the context. We also train and test our models on randomly generated train/dev/test splits followed by an experiment where the models are trained on a select set of genres but tested genres not seen in training. These comprehensive evaluation strategies were performed using three SOTA NER models. Based on our results, we recommend some useful reporting practices for NER researchers, that could help in providing a better understanding of a SOTA model's performance in future.},
  url   = {https://aclanthology.org/2022.lrec-1.643}
}

@InProceedings{takahashi-EtAl:2022:LREC,
  author   = {Takahashi, Yujin and Kaneko, Masahiro and Mita, Masato and Komachi, Mamoru},
  title   = {ProQE: Proficiency-wise Quality Estimation dataset for Grammatical Error Correction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {5994--6000},
  abstract = {This study investigates how supervised quality estimation (QE) models of grammatical error correction (GEC) are affected by the learners' proficiency with the data. QE models for GEC evaluations in prior work have obtained a high correlation with manual evaluations. However, when functioning in a real-world context, the data used for the reported results have limitations because prior works were biased toward data by learners with relatively high proficiency levels. To address this issue, we created a QE dataset that includes multiple proficiency levels and explored the necessity of performing proficiency-wise evaluation for QE of GEC. Our experiments demonstrated that differences in

evaluation dataset proficiency affect the performance of QE models, and proficiency-wise evaluation helps create more robust models.},
  url   = {https://aclanthology.org/2022.lrec-1.644}
}

@InProceedings{tadimeti-georgila-traum:2022:LREC,
  author   = {Tadimeti, Divya and Georgila, Kallirroi and Traum, David},
  title   = {Evaluation of Off-the-shelf Speech Recognizers on Different Accents in a Dialogue Domain},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6001--6008},
  abstract = {We evaluate several publicly available off-the-shelf (commercial and research) automatic speech recognition (ASR) systems on dialogue agent-directed English speech from speakers with General American vs. non-American accents. Our results show that the performance of the ASR systems for non-American accents is considerably worse than for General American accents. Depending on the recognizer, the absolute difference in performance between General American accents and all non-American accents combined can vary approximately from 2\% to 12\%, with relative differences varying approximately between 16\% and 49\%. This drop in performance becomes even larger when we consider specific categories of non-American accents indicating a need for more diligent collection of and training on non-native English speaker data in order to narrow this performance gap. There are performance differences across ASR systems, and while the same general pattern holds, with more errors for non-American accents, there are some accents for which the best recognizer is different than in the overall case. We expect these results to be useful for dialogue system designers in developing more robust inclusive dialogue systems, and for ASR providers in taking into account performance requirements for different accents.},
  url   = {https://aclanthology.org/2022.lrec-1.645}
}

@InProceedings{akula-garibay:2022:LREC,
  author   = {Akula, Ramya and Garibay, Ivan},
  title   = {Sentence Pair Embeddings Based Evaluation Metric for Abstractive and Extractive Summarization},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6009--6017},
  abstract = {The development of an automatic evaluation metric remains an open problem in text generation. Widely used evaluation

metrics, like ROUGE and BLEU, are based on exact word matching and fail to capture semantic similarity. Recent works, such as BERTScore, MoverScore and, Sentence Mover's Similarity, are an improvement over these standard metrics as they use the contextualized word or sentence embeddings to capture semantic similarity. We in this work, propose a novel evaluation metric, Sentence Pair EmbEDdings (SPEED) Score, for text generation which is based on semantic similarity between sentence pairs as opposed to earlier approaches. To find semantic similarity between a pair of sentences, we obtain sentence-level embeddings from multiple transformer models pre-trained specifically on various sentence pair tasks such as Paraphrase Detection (PD), Semantic Text Similarity (STS), and Natural Language Inference (NLI). As these sentence pair tasks involve capturing the semantic similarity between a pair of input texts, we leverage these models in our metric computation. Our proposed evaluation metric shows an impressive performance in evaluating both abstractive and extractive summarization models and achieves state-of-the-art results on the SummEval dataset, demonstrating the effectiveness of our approach. Also, we perform the run-time analysis to show that our proposed metric is faster than the current state-of-the-art.},
  url   = {https://aclanthology.org/2022.lrec-1.646}
}

@InProceedings{poibeau:2022:LREC,
  author   = {Poibeau, Thierry},
  title   = {On ``Human Parity'' and ``Super Human Performance'' in Machine Translation Evaluation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6018--6023},
  abstract = {In this paper, we reassess claims of human parity and super human performance in machine translation. Although these terms have already been discussed, as well as the evaluation protocols used to achieved these conclusions (human-parity is achieved i} only for a very reduced number of languages, ii) on very specific types of documents and iii) with very literal translations), we show that the terms used are themselves problematic, and that human translation involves much more than what is embedded in automatic systems. We also discuss ethical issues related to the way results are presented and advertised. Finally, we claim that a better assessment of human capacities should be put forward and that the goal of replacing humans by machines is not a desirable one.},
  url   = {https://aclanthology.org/2022.lrec-1.647}
}

@InProceedings{araujo-EtAl:2022:LREC,
  author   = {Araujo, Vladimir and Carvallo, Andrés and Kundu, Souvik and Cañete, José and Mendoza, Marcelo and Mercer, Robert E. and Bravo-Marquez, Felipe and Moens, Marie-Francine

and Soto, Alvaro},
  title   = {Evaluation Benchmarks for Spanish Sentence Representations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6024--6034},
  abstract = {Due to the success of pre-trained language models, versions of languages other than English have been released in recent years. This fact implies the need for resources to evaluate these models. In the case of Spanish, there are few ways to systematically assess the models' quality. In this paper, we narrow the gap by building two evaluation benchmarks. Inspired by previous work (Conneau and Kiela, 2018; Chen et al., 2019), we introduce Spanish SentEval and Spanish DiscoEval, aiming to assess the capabilities of stand-alone and discourse-aware sentence representations, respectively. Our benchmarks include considerable pre-existing and newly constructed datasets that address different tasks from various domains. In addition, we evaluate and analyze the most recent pre-trained Spanish language models to exhibit their capabilities and limitations. As an example, we discover that for the case of discourse evaluation tasks, mBERT, a language model trained on multiple languages, usually provides a richer latent representation than models trained only with documents in Spanish. We hope our contribution will motivate a fairer, more comparable, and less cumbersome way to evaluate future Spanish language models.},
  url   = {https://aclanthology.org/2022.lrec-1.648}
}

@InProceedings{garcadaz-EtAl:2022:LREC,
  author   = {García-Díaz, José Antonio and Vivancos-Vicente, Pedro José and Almela, Ángela and Valencia-García, Rafael},
  title   = {UMUTextStats: A linguistic feature extraction tool for Spanish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6035--6044},
  abstract = {Feature Engineering consists in the application of domain knowledge to select and transform relevant features to build efficient machine learning models. In the Natural Language Processing field, the state of the art concerning automatic document classification tasks relies on word and sentence embeddings built upon deep learning models based on transformers that have outperformed the competition in several tasks. However, the models built from these embeddings are usually difficult to interpret. On the contrary, linguistic features are easy to understand, they

result in simpler models, and they usually achieve encouraging results. Moreover, both linguistic features and embeddings can be combined with different strategies which result in more reliable machine-learning models. The de facto tool for extracting linguistic features in Spanish is LIWC. However, this software does not consider specific linguistic phenomena of Spanish such as grammatical gender and lacks certain verb tenses. In order to solve these drawbacks, we have developed UMUTextStats, a linguistic extraction tool designed from scratch for Spanish. Furthermore, this tool has been validated to conduct different experiments in areas such as infodemiology, hate-speech detection, author profiling, authorship verification, humour or irony detection, among others. The results indicate that the combination of linguistic features and embeddings based on transformers are beneficial in automatic document classification.},
  url   = {https://aclanthology.org/2022.lrec-1.649}
}

@InProceedings{prost:2022:LREC,
  author   = {Prost, Jean-Philippe},
  title   = {Integrating a Phrase Structure Corpus Grammar and a Lexical-Semantic Network: the HOLINET Knowledge Graph},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {613--622},
  abstract = {In this paper we address the question of how to integrate grammar and lexical-semantic knowledge within a single and homogeneous knowledge graph. We introduce a graph modelling of grammar knowledge which enables its merging with a lexical-semantic network. Such an integrated representation is expected, for instance, to provide new material for language-related graph embeddings in order to model interactions between Syntax and Semantics. Our base model relies on a phrase structure grammar. The phrase structure is accounted for by both a Proof-Theoretical representation, through a Context-Free Grammar, and a Model-Theoretical one, through a constraint-based grammar. The constraint types colour the grammar layer with syntactic relationships such as Immediate Dominance, Linear Precedence, and more. We detail a creation process which infers the grammar layer from a corpus annotated in constituency and integrates it with a lexical-semantic network through a shared POS tagset. We implement the process, and experiment with the French Treebank and the JeuxDeMots lexical-semantic network. The outcome is the HOLINET knowledge graph.},
  url   = {https://aclanthology.org/2022.lrec-1.65}
}

@InProceedings{heffernan-teufel:2022:LREC,
  author   = {Heffernan, Kevin and Teufel, Simone},
  title   = {Problem-solving Recognition in Scientific Text},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6045--6058},
  abstract = {As far back as Aristotle, problems and solutions have been recognised as a core pattern of thought, and in particular of the scientific method. In this work, we present the novel task of problem-solving recognition in scientific text. Previous work on problem-solving either is not computational, is not adapted to scientific text, or has been narrow in scope. This work provides a new annotation scheme of problem-solving tailored to the scientific domain. We validate the scheme with an annotation study, and model the task using state-of-the-art baselines such as a Neural Relational Topic Model. The agreement study indicates that our annotation is reliable, and results from modelling show that problem-solving expressions in text can be recognised to a high degree of accuracy.},
  url   = {https://aclanthology.org/2022.lrec-1.650}
}

@InProceedings{zhang-yamana:2022:LREC,
  author   = {ZHANG, YUXIANG and Yamana, Hayato},
  title   = {HRCA+: Advanced Multiple-choice Machine Reading Comprehension Method},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6059--6068},
  abstract = {Multiple-choice question answering (MCQA) for machine reading comprehension (MRC) is challenging. It requires a model to select a correct answer from several candidate options related to text passages or dialogue. To select the correct answer, such models must have the ability to understand natural languages, comprehend textual representations, and infer the relationship between candidate options, questions, and passages. Previous models calculated representations between passages and question-option pairs separately, thereby ignoring the effect of other relation-pairs. In this study, we propose a human reading comprehension attention (HRCA) model and a passage-question-option (PQO) matrix-guided HRCA model called HRCA+ to increase accuracy. The HRCA model updates the information learned from the previous relation-pair to the next relation-pair. HRCA+ utilizes the textual information and the interior relationship between every two parts in a passage, a question, and the corresponding candidate options. Our proposed method outperforms other state-of-the-art methods. On the Semeval-2018 Task 11 dataset, our proposed method improved accuracy levels from 95.8\% to 97.2\%, and on the DREAM dataset, it improved accuracy levels from 90.4\% to 91.6\% without extra training data, from 91.8\% to 92.6\% with extra training data.},

  url   = {https://aclanthology.org/2022.lrec-1.651}
}

@InProceedings{parmar-narayan:2022:LREC,
  author   = {Parmar, Maulik and Narayan, Apurva},
  title   = {HyperBox: A Supervised Approach for Hypernym Discovery using Box Embeddings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6069--6076},
  abstract = {Hypernymy plays a fundamental role in many AI tasks like taxonomy learning, ontology learning, etc. This has motivated the development of many automatic identification methods for extracting this relation, most of which rely on word distribution. We present a novel model HyperBox to learn box embeddings for hypernym discovery. Given an input term, HyperBox retrieves its suitable hypernym from a target corpus. For this task, we use the dataset published for SemEval 2018 Shared Task on Hypernym Discovery. We compare the performance of our model on two specific domains of knowledge: medical and music. Experimentally, we show that our model outperforms existing methods on the majority of the evaluation metrics. Moreover, our model generalize well over unseen hypernymy pairs using only a small set of training data.},
  url   = {https://aclanthology.org/2022.lrec-1.652}
}

@InProceedings{xie-EtAl:2022:LREC,
  author   = {Xie, Zhengnan and Kwak, Alice Saebom and George, Enfa and Dozal, Laura W. and Van, Hoang and Jah, Moriba and Furfaro, Roberto and Jansen, Peter},
  title   = {Extracting Space Situational Awareness Events from News Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6077--6082},
  abstract = {Space situational awareness typically makes use of physical measurements from radar, telescopes, and other assets to monitor satellites and other spacecraft for operational, navigational, and defense purposes. In this work we explore using textual input for the space situational awareness task. We construct a corpus of 48.5k news articles spanning all known active satellites between 2009 and 2020. Using a dependency-rule-based extraction system designed to target three high-impact events -- spacecraft launches, failures, and decommissionings, we identify 1,787 space-event sentences that are then annotated by humans with 15.9k labels for event slots. We empirically demonstrate a state-of-the-art

neural extraction system achieves an overall F1 between 53 and 91 per slot for event extraction in this low-resource, high-impact domain.},
  url   = {https://aclanthology.org/2022.lrec-1.653}
}

@InProceedings{jamali-yaghoobzadeh-faili:2022:LREC,
  author   = {Jamali, Naghme and Yaghoobzadeh, Yadollah and Faili, Heshaam},
  title   = {PerCQA: Persian Community Question Answering Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6083--6092},
  abstract = {Community Question Answering (CQA) forums provide answers to many real-life questions. These forums are trendy among machine learning researchers due to their large size. Automatic answer selection, answer ranking, question retrieval, expert finding, and fact-checking are example learning tasks performed using CQA data. This paper presents PerCQA, the first Persian dataset for CQA. This dataset contains the questions and answers crawled from the most well-known Persian forum. After data acquisition, we provide rigorous annotation guidelines in an iterative process and then the annotation of question-answer pairs in SemEvalCQA format. PerCQA contains 989 questions and 21,915 annotated answers. We make PerCQA publicly available to encourage more research in Persian CQA. We also build strong benchmarks for the task of answer selection in PerCQA by using mono- and multi-lingual pre-trained language models.},
  url   = {https://aclanthology.org/2022.lrec-1.654}
}

@InProceedings{lertvittayakumjorn-EtAl:2022:LREC,
  author   = {Lertvittayakumjorn, Piyawat and Choshen, Leshem and Shnarch, Eyal and Toni, Francesca},
  title   = {GrASP: A Library for Extracting and Exploring Human-Interpretable Textual Patterns},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6093--6103},
  abstract = {Data exploration is an important step of every data science and machine learning project, including those involving textual data. We provide a novel language tool, in the form of a publicly available Python library for extracting patterns from textual data. The library integrates a first public implementation of the existing GrASP algorithm. It allows users to extract patterns

using a number of general-purpose built-in linguistic attributes (such as hypernyms, part-of-speech tags, and syntactic dependency tags), as envisaged for the original algorithm, as well as domain-specific custom attributes which can be incorporated into the library by implementing two functions. The library is equipped with a web-based interface empowering human users to conveniently explore data via the extracted patterns, using complementary pattern-centric and example-centric views: the former includes a reading in natural language and statistics of each extracted pattern; the latter shows applications of each extracted pattern to training examples. We demonstrate the usefulness of the library in classification (spam detection and argument mining), model analysis (machine translation), and artifact discovery in datasets (SNLI and 20Newsgroups).},
  url   = {https://aclanthology.org/2022.lrec-1.655}
}

@InProceedings{luo-zhu:2022:LREC,
  author   = {luo, zhaoxin and Zhu, Michael},
  title   = {Recurrent Neural Networks with Mixed Hierarchical Structures and EM Algorithm for Natural Language Processing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6104--6113},
  abstract = {How to obtain hierarchical representations with an increasing level of abstraction becomes one of the key issues of learning with deep neural networks. A variety of RNN models have recently been proposed to incorporate both explicit and implicit hierarchical information in modeling languages in the literature. In this paper, we propose a novel approach called the latent indicator layer to identify and learn implicit hierarchical information (e.g., phrases), and further develop an EM algorithm to handle the latent indicator layer in training. The latent indicator layer further simplifies a text's hierarchical structure, which allows us to seamlessly integrate different levels of attention mechanisms into the structure. We called the resulting architecture as the EM-HRNN model. Furthermore, we develop two bootstrap strategies to effectively and efficiently train the EM-HRNN model on long text documents. Simulation studies and real data applications demonstrate that the EM-HRNN model with bootstrap training outperforms other RNN-based models in document classification tasks. The performance of the EM-HRNN model is comparable to a Transformer-based method called Bert-base, though the former is much smaller model and does not require pre-training.},
  url   = {https://aclanthology.org/2022.lrec-1.656}
}

@InProceedings{jun-EtAl:2022:LREC,
  author   = {Jun, Changwook and Choi, Jooyoung and Sim, Myoseop and Kim, Hyun and Jang, Hansol and Min, Kyungkoo},

  title   = {Korean-Specific Dataset for Table Question Answering},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6114--6120},
  abstract = {Existing question answering systems mainly focus on dealing with text data. However, much of the data produced daily is stored in the form of tables that can be found in documents and relational databases, or on the web. To solve the task of question answering over tables, there exist many datasets for table question answering written in English, but few Korean datasets. In this paper, we demonstrate how we construct Korean-specific datasets for table question answering: Korean tabular dataset is a collection of 1.4M tables with corresponding descriptions for unsupervised pre-training language models. Korean table question answering corpus consists of 70k pairs of questions and answers created by crowd-sourced workers. Subsequently, we then build a pre-trained language model based on Transformer and fine-tune the model for table question answering with these datasets. We then report the evaluation results of our model. We make our datasets publicly available via our GitHub repository and hope that those datasets will help further studies for question answering over tables, and for the transformation of table formats.},
  url   = {https://aclanthology.org/2022.lrec-1.657}
}

@InProceedings{schaefer-stede:2022:LREC,
  author   = {Schaefer, Robin and Stede, Manfred},
  title   = {GerCCT: An Annotated Corpus for Mining Arguments in German Tweets on Climate Change},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6121--6130},
  abstract = {While the field of argument mining has grown notably in the last decade, research on the Twitter medium remains relatively understudied. Given the difficulty of mining arguments in tweets, recent work on creating annotated resources mainly utilized simplified annotation schemes that focus on single argument components, i.e., on claim or evidence. In this paper we strive to fill this research gap by presenting GerCCT, a new corpus of German tweets on climate change, which was annotated for a set of different argument components and properties. Additionally, we labelled sarcasm and toxic language to facilitate the development of tools for filtering out non-argumentative content. This, to the best of our knowledge, renders our corpus the first tweet resource annotated for argumentation, sarcasm and toxic language. We show that a

comparatively complex annotation scheme can still yield promising inter-annotator agreement. We further present first good supervised classification results yielded by a fine-tuned BERT architecture.},
  url   = {https://aclanthology.org/2022.lrec-1.658}
}

@InProceedings{kimura-ototake-sasaki:2022:LREC,
  author   = {Kimura, Yasutomo and Ototake, Hokuto and Sasaki, Minoru},
  title   = {Budget Argument Mining Dataset Using Japanese Minutes from the National Diet and Local Assemblies},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6131--6138},
  abstract = {Budget argument mining attempts to identify argumentative components related to a budget item, and then classifies these argumentative components, given budget information and minutes. We describe the construction of the dataset for budget argument mining, a subtask of QA Lab-PoliInfo-3 in NTCIR-16. Budget argument mining analyses the argument structure of the minutes, focusing on monetary expressions (amount of money). In this task, given sufficient budget information (budget item, budget amount, etc.), relevant argumentative components in the minutes are identified and argument labels (claim, premise, and other) are assigned their components. In this paper, we describe the design of the data format, the annotation procedure, and release information of budget argument mining dataset, to link budget information to minutes.},
  url   = {https://aclanthology.org/2022.lrec-1.659}
}

@InProceedings{ottolina-EtAl:2022:LREC,
  author   = {Ottolina, Giorgio and Palmonari, Matteo Luigi and Vimercati, Manuel and Alam, Mehwish},
  title   = {On the Impact of Temporal Representations on Metaphor Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {623--632},
  abstract = {State-of-the-art approaches for metaphor detection compare their literal - or core - meaning and their contextual meaning using metaphor classifiers based on neural networks. However, metaphorical expressions evolve over time due to various reasons, such as cultural and societal impact. Metaphorical expressions are known to co-evolve with language and literal word meanings, and even drive, to some extent, this evolution. This poses

the question of whether different, possibly time-specific, representations of literal meanings may impact the metaphor detection task. To the best of our knowledge, this is the first study that examines the metaphor detection task with a detailed exploratory analysis where different temporal and static word embeddings are used to account for different representations of literal meanings. Our experimental analysis is based on three popular benchmarks used for metaphor detection and word embeddings extracted from different corpora and temporally aligned using different state-of-the-art approaches. The results suggest that the usage of different static word embedding methods does impact the metaphor detection task and some temporal word embeddings slightly outperform static methods. However, the results also suggest that temporal word embeddings may provide representations of the core meaning of the metaphor even too close to their contextual meaning, thus confusing the classifier. Overall, the interaction between temporal language evolution and metaphor detection appears tiny in the benchmark datasets used in our experiments. This suggests that future work for the computational analysis of this important linguistic phenomenon should first start by creating a new dataset where this interaction is better represented.},
  url   = {https://aclanthology.org/2022.lrec-1.66}
}

@InProceedings{lee-kim-seo:2022:LREC,
  author   = {Lee, Do-Myoung and Kim, Yeachan and Seo, Chang gyun},
  title   = {Context-based Virtual Adversarial Training for Text Classification with Noisy Labels},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6139--6146},
  abstract = {Deep neural networks (DNNs) have a high capacity to completely memorize noisy labels given sufficient training time, and its memorization unfortunately leads to performance degradation. Recently, virtual adversarial training (VAT) attracts attention as it could further improve the generalization of DNNs in semi-supervised learning. The driving force behind VAT is to prevent the models from overffiting to data points by enforcing consistency between the inputs and the perturbed inputs. These strategy could be helpful in learning from noisy labels if it prevents neural models from learning noisy samples while encouraging the models to generalize clean samples. In this paper, we propose context-based virtual adversarial training (ConVAT) to prevent a text classifier from overfitting to noisy labels. Unlike the previous works, the proposed method performs the adversarial training in the context level rather than the inputs. It makes the classifier not only learn its label but also its contextual neighbors, which alleviate the learning from noisy labels by preserving contextual semantics on each data point. We conduct extensive experiments on four text

classification datasets with two types of label noises. Comprehensive experimental results clearly show that the proposed method works quite well even with extremely noisy settings.},
  url   = {https://aclanthology.org/2022.lrec-1.660}
}

@InProceedings{li-ye-zhao:2022:LREC,
  author   = {Li, Chenying and Ye, Wenbo and Zhao, Yilun},
  title   = {FinMath: Injecting a Tree-structured Solver for Question Answering over Financial Reports},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6147--6152},
  abstract = {Answering questions over financial reports containing both tabular and textual data (hybrid data) is challenging as it requires models to select information from financial reports and perform complex quantitative analyses. Although current models have demonstrated a solid capability to solve simple questions, they struggle with complex questions that require a multiple-step numerical reasoning process. This paper proposes a new framework named FinMath, which improves the model's numerical reasoning capacity by injecting a tree-structured neural model to perform multi-step numerical reasoning. Specifically, FinMath extracts supporting evidence from the financial reports given the question in the first phase. In the second phase, a tree-structured neural model is applied to generate a tree expression in a top-down recursive way. Experiments on the TAT-QA dataset show that our proposed approach improves the previous best result by 8.5\% absolute for Exact Match (EM) score (50.1\% to 58.6\%) and 6.1\% absolute for numeracy-focused F1 score (58.0\% to 64.1\%).},
  url   = {https://aclanthology.org/2022.lrec-1.661}
}

@InProceedings{gusev-tikhonov:2022:LREC,
  author   = {Gusev, Ilya and Tikhonov, Alexey},
  title   = {HeadlineCause: A Dataset of News Headlines for Detecting Causalities},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6153--6161},
  abstract = {Detecting implicit causal relations in texts is a task that requires both common sense and world knowledge. Existing datasets are focused either on commonsense causal reasoning or explicit causal relations. In this work, we present HeadlineCause, a dataset for detecting implicit causal relations between pairs of news headlines. The dataset includes over 5000 headline pairs from

English news and over 9000 headline pairs from Russian news labeled through crowdsourcing. The pairs vary from totally unrelated or belonging to the same general topic to the ones including causation and refutation relations. We also present a set of models and experiments that demonstrates the dataset validity, including a multilingual XLM-RoBERTa based model for causality detection and a GPT-2 based model for possible effects prediction.},
  url   = {https://aclanthology.org/2022.lrec-1.662}
}

@InProceedings{liu-EtAl:2022:LREC,
  author   = {Liu, Boyang and Schlegel, Viktor and Batista-Navarro, Riza and Ananiadou, Sophia},
  title   = {Incorporating Zoning Information into Argument Mining from Biomedical Literature},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6162--6169},
  abstract = {The goal of text zoning is to segment a text into zones (i.e., Background, Conclusion) that serve distinct functions. Argumentative zoning, a specific text zoning scheme for the scientific domain, is considered as the antecedent for argument mining by many researchers. Surprisingly, however, little work is concerned with exploiting zoning information to improve the performance of argument mining models, despite the relatedness of the two tasks. In this paper, we propose two transformer-based models to incorporate zoning information into argumentative component identification and classification tasks. One model is for the sentence-level argument mining task and the other is for the token-level task. In particular, we add the zoning labels predicted by an off-the-shelf model to the beginning of each sentence, inspired by the convention commonly used biomedical abstracts. Moreover, we employ multi-head attention to transfer the sentence-level zoning information to each token in a sentence. Based on experiment results, we find a significant improvement in F1-scores for both sentence- and token-level tasks. It is worth mentioning that these zoning labels can be obtained with high accuracy by utilising readily available automated methods. Thus, existing argument mining models can be improved by incorporating zoning information without any additional annotation cost.},
  url   = {https://aclanthology.org/2022.lrec-1.663}
}

@InProceedings{verma-EtAl:2022:LREC,
  author   = {Verma, Yash and Jangra, Anubhav and Saha, Sriparna and Jatowt, Adam and Roy, Dwaipayan},
  title   = {MAKED: Multi-lingual Automatic Keyword Extraction Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6170--6179},
  abstract = {Keyword extraction is an integral task for many downstream problems like clustering, recommendation, search and classification. Development and evaluation of keyword extraction techniques require an exhaustive dataset; however, currently, the community lacks large-scale multi-lingual datasets. In this paper, we present MAKED, a large-scale multi-lingual keyword extraction dataset comprising of 540K+ news articles from British Broadcasting Corporation News (BBC News) spanning 20 languages. It is the first keyword extraction dataset for 11 of these 20 languages. The quality of the dataset is examined by experimentation with several baselines. We believe that the proposed dataset will help advance the field of automatic keyword extraction given its size, diversity in terms of languages used, topics covered and time periods as well as its focus on under-studied languages.},
  url   = {https://aclanthology.org/2022.lrec-1.664}
}

@InProceedings{vacareanu-EtAl:2022:LREC,
  author   = {Vacareanu, Robert and Valenzuela-Escárcega, Marco A. and Gouveia Barbosa, George Caique and Sharp, Rebecca and Hahn-Powell, Gustave and Surdeanu, Mihai},
  title   = {From Examples to Rules: Neural Guided Rule Synthesis for Information Extraction},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6180--6189},
  abstract = {While deep learning approaches to information extraction have had many successes, they can be difficult to augment or maintain as needs shift. Rule-based methods, on the other hand, can be more easily modified. However, crafting rules requires expertise in linguistics and the domain of interest, making it infeasible for most users. Here we attempt to combine the advantages of these two directions while mitigating their drawbacks. We adapt recent advances from the adjacent field of program synthesis to information extraction, synthesizing rules from provided examples. We use a transformer-based architecture to guide an enumerative search, and show that this reduces the number of steps that need to be explored before a rule is found. Further, we show that without training the synthesis algorithm on the specific domain, our synthesized rules achieve state-of-the-art performance on the 1-shot scenario of a task that focuses on few-shot learning for relation classification, and competitive performance in the 5-shot scenario.},
  url   = {https://aclanthology.org/2022.lrec-1.665}
}

@InProceedings{qin-tian-song:2022:LREC,
  author   = {Qin, Han and Tian, Yuanhe and Song, Yan},
  title   = {Enhancing Relation Extraction via Adversarial Multi-task Learning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6190--6199},
  abstract = {Relation extraction (RE) is a sub-field of information extraction, which aims to extract the relation between two given named entities (NEs) in a sentence and thus requires a good understanding of contextual information, especially the entities and their surrounding texts. However, limited attention is paid by most existing studies to re-modeling the given NEs and thus lead to inferior RE results when NEs are sometimes ambiguous. In this paper, we propose a RE model with two training stages, where adversarial multi-task learning is applied to the first training stage to explicitly recover the given NEs so as to enhance the main relation extractor, which is trained alone in the second stage. In doing so, the RE model is optimized by named entity recognition (NER) and thus obtains a detailed understanding of entity-aware context. We further propose the adversarial mechanism to enhance the process, which controls the effect of NER on the main relation extractor and allows the extractor to benefit from NER while keep focusing on RE rather than the entire multi-task learning. Experimental results on two English benchmark datasets for RE demonstrate the effectiveness of our approach, where state-of-the-art performance is observed on both datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.666}
}

@InProceedings{bollegala-machide-kawarabayashi:2022:LREC,
  author   = {Bollegala, Danushka and Machide, Tomoya and Kawarabayashi, Ken-ichi},
  title   = {Query Obfuscation by Semantic Decomposition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6200--6211},
  abstract = {We propose a method to protect the privacy of search engine users by decomposing the queries using semantically \emph{related} and unrelated \emph{distractor} terms. Instead of a single query, the search engine receives multiple decomposed query terms. Next, we reconstruct the search results relevant to the original query term by aggregating the search results retrieved for the decomposed query terms. We show that the word embeddings learnt using a distributed representation learning method can be used to

find semantically related and distractor query terms. We derive the relationship between the \emph{obfuscity} achieved through the proposed query anonymisation method and the \emph{reconstructability} of the original search results using the decomposed queries. We analytically study the risk of discovering the search engine users' information intents under the proposed query obfuscation method, and empirically evaluate its robustness against clustering-based attacks. Our experimental results show that the proposed method can accurately reconstruct the search results for user queries, without compromising the privacy of the search engine users.},
  url   = {https://aclanthology.org/2022.lrec-1.667}
}

@InProceedings{hu-EtAl:2022:LREC,
  author   = {Hu, Ruofan and Zhang, Dongyu and Tao, Dandan and Hartvigsen, Thomas and Feng, Hao and Rundensteiner, Elke},
  title   = {TWEET-FID: An Annotated Dataset for Multiple Foodborne Illness Detection Tasks},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6212--6222},
  abstract = {Foodborne illness is a serious but preventable public health problem — with delays in detecting the associated outbreaks resulting in productivity loss, expensive recalls, public safety hazards, and even loss of life. While social media is a promising source for identifying unreported foodborne illnesses, there is a dearth of labeled datasets for developing effective outbreak detection models. To accelerate the development of machine learning-based models for foodborne outbreak detection, we thus present TWEET-FID (TWEET-Foodborne Illness Detection), the first publicly available annotated dataset for multiple foodborne illness incident detection tasks. TWEET-FID collected from Twitter is annotated with three facets: tweet class, entity type, and slot type, with labels produced by experts as well as by crowdsource workers. We introduce several domain tasks leveraging these three facets: text relevance classification (TRC), entity mention detection (EMD), and slot filling (SF). We describe the end-to-end methodology for dataset design, creation, and labeling for supporting model development for these tasks. A comprehensive set of results for these tasks leveraging state-of-the-art single-and multi-task deep learning methods on the TWEET-FID dataset are provided. This dataset opens opportunities for future research in foodborne outbreak detection.},
  url   = {https://aclanthology.org/2022.lrec-1.668}
}

@InProceedings{skrzewski-pieniowski-demenko:2022:LREC,
  author   = {Skórzewski, Paweł and Pieniowski, Mikołaj and Demenko, Grazyna},
  title   = {Named Entity Recognition to Detect Criminal Texts on

the Web},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6223--6231},
  abstract = {This paper presents a toolkit that applies named-entity extraction techniques to identify information related to criminal activity in texts from the Polish Internet. The methodological and technical assumptions were established following the requirements of our application users from the Border Guard. Due to the specificity of the users' needs and the specificity of web texts, we used original methodologies related to the search for desired texts, the creation of domain lexicons, the annotation of the collected text resources, and the combination of rule-based and machine-learning techniques for extracting the information desired by the user. The performance of our tools has been evaluated on 6240 manually annotated text fragments collected from Internet sources. Evaluation results and user feedback show that our approach is feasible and has potential value for real-life applications in the daily work of border guards. Lexical lookup combined with hand-crafted rules and regular expressions, supported by text statistics, can make a decent specialized entity recognition system in the absence of large data sets required for training a good neural network.},
  url   = {https://aclanthology.org/2022.lrec-1.669}
}

@InProceedings{sileo-moens:2022:LREC,
  author   = {Sileo, Damien and Moens, Marie-Francine},
  title   = {Analysis and Prediction of NLP Models via Task Embeddings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {633--647},
  abstract = {Task embeddings are low-dimensional representations that are trained to capture task properties. In this paper, we propose MetaEval, a collection of 101 NLP tasks. We fit a single transformer to all MetaEval tasks jointly while conditioning it on learned embeddings. The resulting task embeddings enable a novel analysis of the space of tasks. We then show that task aspects can be mapped to task embeddings for new tasks without using any annotated examples. Predicted embeddings can modulate the encoder for zero-shot inference and outperform a zero-shot baseline on GLUE tasks. The provided multitask setup can function as a benchmark for future transfer learning research.},
  url   = {https://aclanthology.org/2022.lrec-1.67}
}

@InProceedings{xu-ouyang-liu:2022:LREC,
  author   = {Xu, zhuoqun and Ouyang, Liubo and Liu, Yang},
  title   = {Task-Driven and Experience-Based Question Answering Corpus for In-Home Robot Application in the House3D Virtual Environment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6232--6239},
  abstract = {At present, more and more work has begun to pay attention to the long-term housekeeping robot scene. Naturally, we wonder whether the robot can answer the questions raised by the owner according to the actual situation at home. These questions usually do not have a clear text context, are directly related to the actual scene, and it is difficult to find the answer from the general knowledge base (such as Wikipedia). Therefore, the experience accumulated from the task seems to be a more natural choice. We present a corpus called TEQA (task-driven and experience-based question answering) in the long-term household task. Based on a popular in-house virtual environment (AI2-THOR) and agent task experiences of ALFRED, we design six types of questions along with answering including 24 question templates, 37 answer templates, and nearly 10k different question answering pairs. Our corpus aims at investigating the ability of task experience understanding of agents for the daily question answering scenario on the ALFRED dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.670}
}

@InProceedings{vanallemeersch-EtAl:2022:LREC,
  author   = {Vanallemeersch, Tom and Defauw, Arne and Szoc, Sara and Kramchaninova, Alina and Van den Bogaert, Joachim and Lösch, Andrea},
  title   = {ELRC Action: Covering Confidentiality, Correctness and Cross-linguality},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6240--6249},
  abstract = {We describe the language technology (LT) assessments carried out in the ELRC action (European Language Resource Coordination) of the European Commission, which aims towards minimising language barriers across the EU. We zoom in on the two most extensive assessments. These LT specifications do not only involve experiments with tools and techniques but also an extensive consultation round with stakeholders from public organisations, academia and industry, in order to gather insights into scenarios and best practices. The LT specifications concern (1) the field of

automated anonymisation, which is motivated by the need of public and other organisations to be able to store and share data, and (2) the field of multilingual fake news processing, which is motivated by the increasingly pressing problem of disinformation and the limited language coverage of systems for automatically detecting misleading articles. For each specification, we set up a corresponding proof-of-concept software to demonstrate the opportunities and challenges involved in the field.},
  url   = {https://aclanthology.org/2022.lrec-1.671}
}

@InProceedings{soni-EtAl:2022:LREC,
  author   = {Soni, Sarvesh and Gudala, Meghana and Pajouhi, Atieh and Roberts, Kirk},
  title   = {RadQA: A Question Answering Dataset to Improve Comprehension of Radiology Reports},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6250--6259},
  abstract = {We present a radiology question answering dataset, RadQA, with 3074 questions posed against radiology reports and annotated with their corresponding answer spans (resulting in a total of 6148 question-answer evidence pairs) by physicians. The questions are manually created using the clinical referral section of the reports that take into account the actual information needs of ordering physicians and eliminate bias from seeing the answer context (and, further, organically create unanswerable questions). The answer spans are marked within the Findings and Impressions sections of a report. The dataset aims to satisfy the complex clinical requirements by including complete (yet concise) answer phrases (which are not just entities) that can span multiple lines. We conduct a thorough analysis of the proposed dataset by examining the broad categories of disagreement in annotation (providing insights on the errors made by humans) and the reasoning requirements to answer a question (uncovering the huge dependence on medical knowledge for answering the questions). The advanced transformer language models achieve the best F1 score of 63.55 on the test set, however, the best human performance is 90.31 (with an average of 84.52). This demonstrates the challenging nature of RadQA that leaves ample scope for future method research.},
  url   = {https://aclanthology.org/2022.lrec-1.672}
}

@InProceedings{agarwal-EtAl:2022:LREC,
  author   = {Agarwal, Ankush and Gite, Raj and Laddha, Shreya and Bhattacharyya, Pushpak and Kar, Satyanarayan and Ekbal, Asif and Thind, Prabhjit and Zele, Rajesh and Shankar, Ravi},
  title   = {Knowledge Graph - Deep Learning: A Case Study in Question Answering in Aviation Safety Domain},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6260--6270},
  abstract = {In the commercial aviation domain, there are a large number of documents, like accident reports of NTSB and ASRS, and regulatory directives ADs. There is a need for a system to efficiently access these diverse repositories to serve the demands of the aviation industry, such as maintenance, compliance, and safety. In this paper, we propose a Knowledge Graph (KG) guided Deep Learning (DL) based Question Answering (QA) system to cater to these requirements. We construct a KG from aircraft accident reports and contribute this resource to the community of researchers. The efficacy of this resource is tested and proved by the proposed QA system. Questions in Natural Language are converted into SPARQL (the interface language of the RDF graph database) queries and are answered from the KG. On the DL side, we examine two different QA models, BERT-QA and GPT3-QA, covering the two paradigms of answer formulation in QA. We evaluate our system on a set of handcrafted queries curated from the accident reports. Our hybrid KG + DL QA system, KGQA + BERT-QA, achieves 7\% and 40.3\% increase in accuracy over KGQA and BERT-QA systems respectively. Similarly, the other combined system, KGQA + GPT3-QA, achieves 29.3\% and 9.3\% increase in accuracy over KGQA and GPT3-QA systems respectively. Thus, we infer that the combination of KG and DL is better than either KG or DL individually for QA, at least in our chosen domain.},
  url   = {https://aclanthology.org/2022.lrec-1.673}
}

@InProceedings{wood-arnold-wang:2022:LREC,
  author   = {Wood, Justin and Arnold, Corey and Wang, Wei},
  title   = {A Bayesian Topic Model for Human-Evaluated Interpretability},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6271--6279},
  abstract = {One desiderata of topic modeling is to produce interpretable topics. Given a cluster of document-tokens comprising a topic, we can order the topic by counting each word. It is natural to think that each topic could easily be labeled by looking at the words with the highest word count. However, this is not always the case. A human evaluator can often have difficulty identifying a single label that accurately describes the topic as many top words seem unrelated. This paper aims to improve interpretability in topic modeling by providing a novel, outperforming interpretable topic model Our approach combines two previously established subdomains in topic modeling: nonparametric and weakly-supervised topic models. Given a nonparametric topic model, we can include weakly-supervised

input using novel modifications to the nonparametric generative model. These modifications lay the groundwork for a compelling setting---one in which most corpora, without any previous supervised or weakly-supervised input, can discover interpretable topics. This setting also presents various challenging sub-problems of which we provide resolutions. Combining nonparametric topic models with weakly-supervised topic models leads to an exciting discovery---a complete, self-contained and outperforming topic model for interpretability.},
  url   = {https://aclanthology.org/2022.lrec-1.674}
}

@InProceedings{faralli-lenzi-velardi:2022:LREC,
  author   = {Faralli, Stefano and Lenzi, Andrea and Velardi, Paola},
  title   = {A Large Interlinked Knowledge Graph of the Italian Cultural Heritage},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6280--6289},
  abstract = {Knowledge is the lifeblood for a plethora of applications such as search, recommender systems and natural language understanding. Thanks to the efforts in the fields of Semantic Web and Linked Open Data a growing number of interlinked knowledge bases are supporting the development of advanced knowledge-based applications. Unfortunately, for a large number of domain-specific applications, these knowledge bases are unavailable. In this paper, we present a resource consisting of a large knowledge graph linking the Italian cultural heritage entities (defined in the ArCo ontology) with the concepts defined on well-known knowledge bases (i.e., DBpedia and the Getty GVP ontology). We describe the methodologies adopted for the semi-automatic resource creation and provide an in-depth analysis of the resulting interlinked graph.},
  url   = {https://aclanthology.org/2022.lrec-1.675}
}

@InProceedings{church-cai-bian:2022:LREC,
  author   = {Church, Kenneth and Cai, Xingyu and Bian, Yuchen},
  title   = {Training on Lexical Resources},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6290--6299},
  abstract = {We propose using lexical resources (thesaurus, VAD) to fine-tune pretrained deep nets such as BERT and ERNIE. Then at inference time, these nets can be used to distinguish synonyms from

antonyms, as well as VAD distances. The inference method can be applied to words as well as texts such as multiword expressions (MWEs), out of vocabulary words (OOVs), morphological variants and more. Code and data are posted on https://github.com/kwchurch/syn_ant.},
  url   = {https://aclanthology.org/2022.lrec-1.676}
}

@InProceedings{cornell-EtAl:2022:LREC,
  author   = {Cornell, Filip and zhang, Chenda and Karlgren, Jussi and Girdzijauskas, Sarunas},
  title   = {Challenging the Assumption of Structure-based embeddings in Few- and Zero-shot Knowledge Graph Completion},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6300--6309},
  abstract = {In this paper, we report experiments on Few- and Zero-shot Knowledge Graph completion, where the objective is to add missing relational links between entities into an existing Knowledge Graph with few or no previous examples of the relation in question. While previous work has used pre-trained embeddings based on the structure of the graph as input for a neural network, nobody has, to the best of our knowledge, addressed the task by only using textual descriptive data associated with the entities and relations, much since current standard benchmark data sets lack such information. We therefore enrich the benchmark data sets for these tasks by collecting textual description data to provide a new resource for future research to bridge the gap between structural and textual Knowledge Graph completion. Our results show that we can improve the results for Knowledge Graph completion for both Few- and Zero-shot scenarios with up to a two-fold increase of all metrics in the Zero-shot setting. From a more general perspective, our experiments demonstrate the value of using textual resources to enrich more formal representations of human knowledge and in the utility of transfer learning from textual data and text collections to enrich and maintain knowledge resources.},
  url   = {https://aclanthology.org/2022.lrec-1.677}
}

@InProceedings{lagzdi-EtAl:2022:LREC,
  author   = {Lagzdiņš, Andis and Siliņš, Uldis and Bergmanis, Toms and Pinnis, Mārcis and Vasiļevskis, Artūrs and Vasiļjevs, Andrejs},
  title   = {Open Terminology Management and Sharing Toolkit for Federation of Terminology Databases},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {6310--6316},
  abstract = {Consolidated access to current and reliable terms from different subject fields and languages is necessary for content creators and translators. Terminology is also needed in AI applications such as machine translation, speech recognition, information extraction, and other natural language processing tools.In this work, we facilitate standards-based sharing and management of terminology resources by providing an open terminology management solution - the EuroTermBank Toolkit. It allows organisations to manage and search their terms, create term collections, and share them within and outside the organisation by participating in the network of federated databases. The data curated in the federated databases are automatically shared with EuroTermBank, the largest multilingual terminology resource in Europe, allowing translators and language service providers as well as researchers and students to access terminology resources in their most current version.},
  url   = {https://aclanthology.org/2022.lrec-1.678}
}

@InProceedings{schoene-dethlefs-ananiadou:2022:LREC,
  author   = {Schoene, Annika Marie and Dethlefs, Nina and Ananiadou, Sophia},
  title   = {RELATE: Generating a linguistically inspired Knowledge Graph for fine-grained emotion classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6317--6327},
  abstract = {Several existing resources are available for sentiment analysis (SA) tasks that are used for learning sentiment specific embedding (SSE) representations. These resources are either large, common-sense knowledge graphs (KG) that cover a limited amount of polarities/emotions or they are smaller in size (e.g.: lexicons), which require costly human annotation and cover fine-grained emotions. Therefore using knowledge resources to learn SSE representations is either limited by the low coverage of polarities/emotions or the overall size of a resource. In this paper, we first introduce a new directed KG called ‘RELATE’, which is built to overcome both the issue of low coverage of emotions and the issue of scalability. RELATE is the first KG of its size to cover Ekman’s six basic emotions that are directed towards entities. It is based on linguistic rules to incorporate the benefit of semantics without relying on costly human annotation. The performance of ‘RELATE’ is evaluated by learning SSE representations using a Graph Convolutional Neural Network (GCN).},
  url   = {https://aclanthology.org/2022.lrec-1.679}
}

@InProceedings{hazem-EtAl:2022:LREC,

  author   = {Hazem, Amir and Bouhandi, Merieme and Boudin, Florian and Daille, Beatrice},
  title   = {Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {648--662},
  abstract = {Automatic Term Extraction (ATE) is a key component for domain knowledge understanding and an important basis for further natural language processing applications. Even with persistent improvements, ATE still exhibits weak results exacerbated by small training data inherent to specialized domain corpora. Recently, transformers-based deep neural models, such as BERT, have proven to be efficient in many downstream NLP tasks. However, no systematic evaluation of ATE has been conducted so far. In this paper, we run an extensive study on fine-tuning pre-trained BERT models for ATE. We propose strategies that empirically show BERT's effectiveness using cross-lingual and cross-domain transfer learning to extract single and multi-word terms. Experiments have been conducted on four specialized domains in three languages. The obtained results suggest that BERT can capture cross-domain and cross-lingual terminologically-marked contexts shared by terms, opening a new design-pattern for ATE.},
  url   = {https://aclanthology.org/2022.lrec-1.68}
}

@InProceedings{markl-mcnulty:2022:LREC,
  author   = {Markl, Nina and McNulty, Stephen Joseph},
  title   = {Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6328--6339},
  abstract = {Despite the fact that variation is a fundamental characteristic of natural language, automatic speech recognition systems perform systematically worse on non-standardised and marginalised language varieties. In this paper we use the lens of language policy to analyse how current practices in training and testing ASR systems in industry lead to the data bias giving rise to these systematic error differences. We believe that this is a useful perspective for speech and language technology practitioners to understand the origins and harms of algorithmic bias, and how they can mitigate it. We also propose a re-framing of language resources as (public) infrastructure which should not solely be designed for markets, but for, and with meaningful cooperation of, speech communities.},

  url   = {https://aclanthology.org/2022.lrec-1.680}
}

@InProceedings{alyafeai-EtAl:2022:LREC,
  author   = {Alyafeai, Zaid and Masoud, Maraim and Ghaleb, Mustafa and Al-shaibani, Maged S.},
  title   = {Masader: Metadata Sourcing for Arabic Text and Speech Data Resources},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6340--6351},
  abstract = {The NLP pipeline has evolved dramatically in the last few years. The first step in the pipeline is to find suitable annotated datasets to evaluate the tasks we are trying to solve. Unfortunately, most of the published datasets lack metadata annotations that describe their attributes. Not to mention, the absence of a public catalogue that indexes all the publicly available datasets related to specific regions or languages. When we consider low-resource dialectical languages, for example, this issue becomes more prominent. In this paper, we create Masader, the largest public catalogue for Arabic NLP datasets, which consists of 200 datasets annotated with 25 attributes. Furthermore, we develop a metadata annotation strategy that could be extended to other languages. We also make remarks and highlight some issues about the current status of Arabic NLP datasets and suggest recommendations to address them.},
  url   = {https://aclanthology.org/2022.lrec-1.681}
}

@InProceedings{robin-EtAl:2022:LREC,
  author   = {Robin, Cécile and Suresh, Gautham Vadakkekara and Rodriguez-Doncel, Víctor and McCrae, John P. and Buitelaar, Paul},
  title   = {Linghub2: Language Resource Discovery Tool for Language Technologies},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6352--6360},
  abstract = {Language resources are a key component of natural language processing and related research and applications. Users of language resources have different needs in terms of format, language, topics, etc. for the data they need to use. Linghub (McCrae and Cimiano, 2015) was first developed for this purpose, using the capabilities of linked data to represent metadata, and tackling the heterogeneous metadata issue. Linghub aimed at helping language resources and technology users to easily find and retrieve

relevant data, and identify important information on access, topics, etc. This work describes a rejuvenation and modernisation of the 2015 platform into using a popular open source data management system, DSpace, as foundation. The new platform, Linghub2, contains updated and extended resources, more languages offered, and continues the work towards homogenisation of metadata through conversions, through linkage to standardisation strategies and community groups, such as the Open Digital Rights Language (ODRL) community group.},
  url   = {https://aclanthology.org/2022.lrec-1.682}
}

@InProceedings{tseng-EtAl:2022:LREC,
  author   = {Tseng, Yu-Hsiang and Shih, Cing-Fang and Chen, Pin-Er and Chou, Hsin-Yu and Ku, Mao-Chang and HSIEH, Shu-Kai},
  title   = {CxLM: A Construction and Context-aware Language Model},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6361--6369},
  abstract = {Constructions are direct form-meaning pairs with possible schematic slots. These slots are simultaneously constrained by the embedded construction itself and the sentential context. We propose that the constraint could be described by a conditional probability distribution. However, as this conditional probability is inevitably complex, we utilize language models to capture this distribution. Therefore, we build CxLM, a deep learning-based masked language model explicitly tuned to constructions' schematic slots. We first compile a construction dataset consisting of over ten thousand constructions in Taiwan Mandarin. Next, an experiment is conducted on the dataset to examine to what extent a pretrained masked language model is aware of the constructions. We then fine-tune the model specifically to perform a cloze task on the opening slots. We find that the fine-tuned model predicts masked slots more accurately than baselines and generates both structurally and semantically plausible word samples. Finally, we release CxLM and its dataset as publicly available resources and hope to serve as new quantitative tools in studying construction grammar.},
  url   = {https://aclanthology.org/2022.lrec-1.683}
}

@InProceedings{hai-EtAl:2022:LREC,
  author   = {Hai, Oufan and Sundberg, Matthew and Trice, Katherine and Friedman, Rebecca and Grimm, Scott},
  title   = {The Lexometer: A Shiny Application for Exploratory Analysis and Visualization of Corpus Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6370--6376},
  abstract = {Often performing even simple data science tasks with corpus data requires significant expertise in data science and programming languages like R and Python. With the aim of making quantitative research more accessible for researchers in the language sciences, we present the Lexometer, a Shiny application that integrates numerous data analysis and visualization functions into an easy-to-use graphical user interface. Some functions of the Lexometer are: filtering large databases to generate subsets of the data and variables of interest, providing a range of graphing techniques for both single and multiple variable analysis, and providing the data in a table format which can further be filtered as well as provide methods for cleaning the data. The Lexometer aims to be useful to language researchers with differing levels of programming expertise and to aid in broadening the inclusion of corpus-based empirical evidence in the language sciences.},
  url   = {https://aclanthology.org/2022.lrec-1.684}
}

@InProceedings{robertson-chang-syrinki:2022:LREC,
  author   = {Robertson, Frankie and Chang, Li-Hsin and Söyrinki, Sini},
  title   = {TallVocabL2Fi: A Tall Dataset of 15 Finnish L2 Learners' Vocabulary},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6377--6386},
  abstract = {Previous work concerning measurement of second language learners has tended to focus on the knowledge of small numbers of words, often geared towards measuring vocabulary size. This paper presents a "tall" dataset containing information about a few learners' knowledge of many words, suitable for evaluating Vocabulary Inventory Prediction (VIP) techniques, including those based on Computerised Adaptive Testing (CAT). In comparison to previous comparable datasets, the learners are from varied backgrounds, so as to reduce the risk of overfitting when used for machine learning based VIP. The dataset contains both a self-rating test and a translation test, used to derive a measure of reliability for learner responses. The dataset creation process is documented, and the relationship between variables concerning the participants, such as their completion time, their language ability level, and the triangulated reliability of their self-assessment responses, are analysed. The word list is constructed by taking into account the extensive derivation morphology of Finnish, and infrequent words are included in order to account for explanatory variables beyond word frequency.},
  url   = {https://aclanthology.org/2022.lrec-1.685}

}

@InProceedings{garg-EtAl:2022:LREC1,
  author   = {Garg, Muskan and Saxena, Chandni and Saha, Sriparna and Krishnan, Veena and Joshi, Ruchi and Mago, Vijay},
  title   = {CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6387--6396},
  abstract = {The social NLP researchers and mental health practitioners have witnessed exponential growth in the field of mental health detection and analysis on social media. It has become important to identify the reason behind mental illness. In this context, we introduce a new dataset for Causal Analysis of Mental health in Social media posts (CAMS). We first introduce the annotation schema for this task of causal analysis. The causal analysis comprises of two types of annotations, viz, causal interpretation and causal categorization. We show the efficacy of our scheme in two ways: (i) crawling and annotating 3155 Reddit data and (ii) re-annotate the publicly available SDCNL dataset of 1896 instances for interpretable causal analysis. We further combine them as CAMS dataset and make it available along with the other source codes {https://anonymous.4open.science/r/CAMS1/}. Our experimental results show that the hybrid CNN-LSTM model gives the best performance over CAMS dataset.},
  url   = {https://aclanthology.org/2022.lrec-1.686}
}

@InProceedings{zhang-wang-zong:2022:LREC,
  author   = {Zhang, Xiaohan and Wang, Shaonan and Zong, Chengqing},
  title   = {How Does the Experimental Setting Affect the Conclusions of Neural Encoding Models?},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6397--6404},
  abstract = {Recent years have witnessed the tendency of neural encoding models on exploring brain language processing using naturalistic stimuli. Neural encoding models are data-driven methods that require an encoding model to investigate the mystery of brain mechanisms hidden in the data. As a data-driven method, the performance of encoding models is very sensitive to the experimental setting. However, it is unknown how the experimental setting further affects the conclusions of neural encoding models. This paper

systematically investigated this problem and evaluated the influence of three experimental settings, i.e., the data size, the cross-validation training method, and the statistical testing method. Results demonstrate that inappropriate cross-validation training and small data size can substantially decrease the performance of encoding models, especially in the temporal lobe and the frontal lobe. And different null hypotheses in significance testing lead to highly different significant brain regions. Based on these results, we suggest a block-wise cross-validation training method and an adequate data size for increasing the performance of linear encoding models. We also propose two strict null hypotheses to control false positive discovery rates.},
  url   = {https://aclanthology.org/2022.lrec-1.687}
}

@InProceedings{kerz-EtAl:2022:LREC,
  author   = {Kerz, Elma and Qiao, Yu and Zanwar, Sourabh and Wiechmann, Daniel},
  title   = {SPADE: A Big Five-Mturk Dataset of Argumentative Speech Enriched with Socio-Demographics for Personality Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6405--6419},
  abstract = {In recent years, there has been increasing interest in automatic personality detection based on language. Progress in this area is highly contingent upon the availability of datasets and benchmark corpora. However, publicly available datasets for modeling and predicting personality traits are still scarce. While recent efforts to create such datasets from social media (Twitter, Reddit) are to be applauded, they often do not include continuous and contextualized language use. In this paper, we introduce SPADE, the first dataset with continuous samples of argumentative speech labeled with the Big Five personality traits and enriched with socio-demographic data (age, gender, education level, language background). We provide benchmark models for this dataset to facilitate further research and conduct extensive experiments. Our models leverage 436 (psycho)linguistic features extracted from transcribed speech and speaker-level metainformation with transformers. We conduct feature ablation experiments to investigate which types of features contribute to the prediction of individual personality traits.},
  url   = {https://aclanthology.org/2022.lrec-1.688}
}

@InProceedings{gupta-boulianne:2022:LREC,
  author   = {gupta, vishwa and Boulianne, Gilles},
  title   = {Progress in Multilingual Speech Recognition for Low Resource Languages Kurmanji Kurdish, Cree and Inuktut},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6420--6428},
  abstract = {This contribution presents our efforts to develop the automatic speech recognition (ASR) systems for three low resource languages: Kurmanji Kurdish, Cree and Inuktut. As a first step, we generate multilingual models from acoustic training data from 12 different languages in the hybrid DNN/HMM framework. We explore different strategies for combining the phones from different languages: either keep the phone labels separate for each language or merge the common phones. For Kurmanji Kurdish and Inuktut, keeping the phones separate gives much lower word error rate (WER), while merging phones gives lower WER for Cree. These WER are lower than training the acoustic models separately for each language. We also compare two different DNN architectures: factored time delay neural network (TDNN-F), and bidirectional long short-term memory (BLSTM) acoustic models. The TDNN-F acoustic models give significantly lower WER for Kurmanji Kurdish and Cree, while BLSTM acoustic models give significantly lower WER for Inuktut. We also show that for each language, training multilingual acoustic models by one more epoch with acoustic data from that language reduces the WER significantly. We also added 512-dimensional embedding features from cross-lingual pre-trained wav2vec2.0 XLSR-53 models, but they lead to only a small reduction in WER.},
  url   = {https://aclanthology.org/2022.lrec-1.689}
}

@InProceedings{jurkschat-EtAl:2022:LREC,
  author   = {Jurkschat, Lena and Wiedemann, Gregor and Heinrich, Maximilian and Ruckdeschel, Mattes and Torge, Sunna},
  title   = {Few-Shot Learning for Argument Aspects of the Nuclear Energy Debate},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {663--672},
  abstract = {We approach aspect-based argument mining as a supervised machine learning task to classify arguments into semantically coherent groups referring to the same defined aspect categories. As an exemplary use case, we introduce the Argument Aspect Corpus - Nuclear Energy that separates arguments about the topic of nuclear energy into nine major aspects. Since the collection of training data for further aspects and topics is costly, we investigate the potential for current transformer-based few-shot learning approaches to accurately classify argument aspects. The best approach is applied to a British newspaper corpus covering the debate on nuclear energy over the past 21 years. Our evaluation shows that a stable prediction of shares of argument aspects in this debate is feasible with 50 to 100 training samples

per aspect. Moreover, we see signals for a clear shift in the public discourse in favor of nuclear energy in recent years. This revelation of changing patterns of pro and contra arguments related to certain aspects over time demonstrates the potential of supervised argument aspect detection for tracking issue-specific media discourses.},
  url   = {https://aclanthology.org/2022.lrec-1.69}
}

@InProceedings{garciaduran-arora-west:2022:LREC,
  author   = {Garcia-Duran, Alberto and Arora, Akhil and West, Robert},
  title   = {Efficient Entity Candidate Generation for Low-Resource Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6429--6438},
  abstract = {Candidate generation is a crucial module in entity linking. It also plays a key role in multiple NLP tasks that have been proven to beneficially leverage knowledge bases. Nevertheless, it has often been overlooked in the monolingual English entity linking literature, as naïve approaches obtain very good performance. Unfortunately, the existing approaches for English cannot be successfully transferred to poorly resourced languages. This paper constitutes an in-depth analysis of the candidate generation problem in the context of cross-lingual entity linking with a focus on low-resource languages. Among other contributions, we point out limitations in the evaluation conducted in previous works. We introduce a characterization of queries into types based on their difficulty, which improves the interpretability of the performance of different methods. We also propose a light-weight and simple solution based on the construction of indexes whose design is motivated by more complex transfer learning based neural approaches. A thorough empirical analysis on 9 real-world datasets under 2 evaluation settings shows that our simple solution outperforms the state-of-the-art approach in terms of both quality and efficiency for almost all datasets and query types.},
  url   = {https://aclanthology.org/2022.lrec-1.690}
}

@InProceedings{lent-EtAl:2022:LREC,
  author   = {Lent, Heather and Ogueji, Kelechi and de Lhoneux, Miryam and Ahia, Orevaoghene and Søgaard, Anders},
  title   = {What a Creole Wants, What a Creole Needs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {6439--6449},
  abstract = {In recent years, the natural language processing (NLP) community has given increased attention to the disparity of efforts directed towards high-resource languages over low-resource ones. Efforts to remedy this delta often begin with translations of existing English datasets into other languages. However, this approach ignores that different language communities have different needs. We consider a group of low-resource languages, creole languages. Creoles are both largely absent from the NLP literature, and also often ignored by society at large due to stigma, despite these languages having sizable and vibrant communities. We demonstrate, through conversations with creole experts and surveys of creole-speaking communities, how the things needed from language technology can change dramatically from one language to another, even when the languages are considered to be very similar to each other, as with creoles. We discuss the prominent themes arising from these conversations, and ultimately demonstrate that useful language technology cannot be built without involving the relevant community.},
  url   = {https://aclanthology.org/2022.lrec-1.691}
}

@InProceedings{gutkin-EtAl:2022:LREC,
  author   = {Gutkin, Alexander and Johny, Cibu and Doctor, Raiomond and Wolf-Sonkin, Lawrence and Roark, Brian},
  title   = {Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6450--6460},
  abstract = {The Brahmic family of scripts is used to record some of the most spoken languages in the world and is arguably the most diverse family of writing systems. In this work, we present several substantial extensions to Brahmic script functionality within the open-source Nisaba library of finite-state script normalization and processing utilities (Johny et al., 2021). First, we extend coverage from the original ten scripts to an additional ten scripts of South Asia and beyond, including some used to record endangered languages such as Dogri. Second, we augment the language layer so that scripts used by multiple languages in distinct ways can be processed correctly for more languages, such as the Bengali script when used for the low-resource language Santali. We document key changes to the finite-state engine required to support these new languages and scripts. Finally, we add new script processing utilities, including lightweight script-level reading normalization that (unlike existing visual normalization) does not preserve visual invariance, and a fixed-input transliteration mechanism specifically tailored to Brahmic text entry with ASCII characters.},
  url   = {https://aclanthology.org/2022.lrec-1.692}
}

@InProceedings{dunn-li-sastre:2022:LREC,
  author   = {Dunn, Jonathan and Li, Haipeng and Sastre, Damian},
  title   = {Predicting Embedding Reliability in Low-Resource Settings Using Corpus Similarity Measures},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6461--6470},
  abstract = {This paper simulates a low-resource setting across 17 languages in order to evaluate embedding similarity, stability, and reliability under different conditions. The goal is to use corpus similarity measures before training to predict properties of embeddings after training. The main contribution of the paper is to show that it is possible to predict downstream embedding similarity using upstream corpus similarity measures. This finding is then applied to low-resource settings by modelling the reliability of embeddings created from very limited training data. Results show that it is possible to estimate the reliability of low-resource embeddings using corpus similarity measures that remain robust on small amounts of data. These findings have significant implications for the evaluation of truly low-resource languages in which such systematic downstream validation methods are not possible because of data limitations.},
  url   = {https://aclanthology.org/2022.lrec-1.693}
}

@InProceedings{abdulmumin-EtAl:2022:LREC,
  author   = {Abdulmumin, Idris and Dash, Satya Ranjan and Dawud, Musa Abdullahi and Parida, Shantipriya and Muhammad, Shamsuddeen and Ahmad, Ibrahim Sa'id and Panda, Subhadarshi and Bojar, Ondřej and Galadanci, Bashir Shehu and Bello, Bello Shehu},
  title   = {Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6471--6479},
  abstract = {Multi-modal Machine Translation (MMT) enables the use of visual information to enhance the quality of translations, especially where the full context is not available to enable the unambiguous translation in standard machine translation. Despite the increasing popularity of such technique, it lacks sufficient and qualitative datasets to maximize the full extent of its potential. Hausa, a Chadic language, is a member of the Afro-Asiatic language family. It is estimated that about 100 to 150 million people speak

the language, with more than 80 million indigenous speakers. This is more than any of the other Chadic languages. Despite the large number of speakers, the Hausa language is considered as a low resource language in natural language processing (NLP). This is due to the absence of enough resources to implement most of the tasks in NLP. While some datasets exist, they are either scarce, machine-generated or in the religious domain. Therefore, there is the need to create training and evaluation data for implementing machine learning tasks and bridging the research gap in the language. This work presents the Hausa Visual Genome (HaVG), a dataset that contains the description of an image or a section within the image in Hausa and its equivalent in English. The dataset was prepared by automatically translating the English description of the images in the Hindi Visual Genome (HVG). The synthetic Hausa data was then carefully postedited, taking into cognizance the respective images. The data is made of 32,923 images and their descriptions that are divided into training, development, test, and challenge test set. The Hausa Visual Genome is the first dataset of its kind and can be used for Hausa-English machine translation, multi-modal research, image description, among various other natural language processing and generation tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.694}
}

@InProceedings{nwafor-andy:2022:LREC,
  author   = {Nwafor, Ebelechukwu and Andy, Anietie},
  title   = {A Survey of Machine Translation Tasks on Nigerian Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6480--6486},
  abstract = {Machine translation is an active area of research that has received a significant amount of attention over the past decade. With the advent of deep learning models, the translation of several languages has been performed with high accuracy and precision. In spite of the development in machine translation techniques, there is very limited work focused on translating low-resource African languages, particularly Nigerian languages. Nigeria is one of the most populous countries in Africa with diverse language and ethnic groups. In this paper, we survey the current state of the art of machine translation research on Nigerian languages with a major emphasis on neural machine translation techniques. We outline the limitations of research in machine translation on Nigerian languages and propose future directions in increasing research and participation.},
  url   = {https://aclanthology.org/2022.lrec-1.695}
}

@InProceedings{yu-EtAl:2022:LREC3,
  author   = {Yu, Tiezheng and Frieske, Rita and Xu, Peng and

Cahyawijaya, Samuel and YIU, Cheuk Tung and Lovenia, Holy and Dai, Wenliang and Barezi, Elham J. and Chen, Qifeng and Ma, Xiaojuan and Shi, Bertram and Fung, Pascale},
  title = {Automatic Speech Recognition Datasets in Cantonese: A Survey and New Dataset},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {6487--6494},
  abstract = {Automatic speech recognition (ASR) on low resource languages improves the access of linguistic minorities to technological advantages provided by artificial intelligence (AI). In this paper, we address the problem of data scarcity for the Hong Kong Cantonese language by creating a new Cantonese dataset. Our dataset, Multi-Domain Cantonese Corpus (MDCC), consists of 73.6 hours of clean read speech paired with transcripts, collected from Cantonese audiobooks from Hong Kong. It comprises philosophy, politics, education, culture, lifestyle and family domains, covering a wide range of topics. We also review all existing Cantonese datasets and analyze them according to their speech type, data source, total size and availability. We further conduct experiments with Fairseq S2T Transformer, a state-of-the-art ASR model, on the biggest existing dataset, Common Voice zh-HK, and our proposed MDCC, and the results show the effectiveness of our dataset. In addition, we create a powerful and robust Cantonese ASR model by applying multi-dataset learning on MDCC and Common Voice zh-HK.},
  url = {https://aclanthology.org/2022.lrec-1.696}
}

@InProceedings{arreerard-mander-piao:2022:LREC,
  author = {Arreerard, Ratchakrit and Mander, Stephen and Piao, Scott},
  title = {Survey on Thai NLP Language Resources and Tools},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {6495--6505},
  abstract = {Over the past decades, Natural Language Processing (NLP) research has been expanding to cover more languages. Recently particularly, NLP community has paid increasing attention to under-resourced languages. However, there are still many languages for which NLP research is limited in terms of both language resources and software tools. Thai language is one of the under-resourced languages in the NLP domain, although it is spoken by nearly 70 million people globally. In this paper, we report on our survey on the past development of Thai NLP research to help understand its current state and future research directions. Our survey shows that, although Thai NLP community has achieved a significant achievement

over the past three decades, particularly on NLP upstream tasks such as tokenisation, research on downstream tasks such as syntactic parsing and semantic analysis is still limited. But we foresee that Thai NLP research will advance rapidly as richer Thai language resources and more robust NLP techniques become available.},
  url   = {https://aclanthology.org/2022.lrec-1.697}
}

@InProceedings{lin-EtAl:2022:LREC,
  author   = {Lin, Nankai and Fu, Yingwen and Chen, Chuwei and Yang, Ziyu and JIANG, Shengyi},
  title   = {LaoPLM: Pre-trained Language Models for Lao},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6506--6512},
  abstract = {Trained on the large corpus, pre-trained language models (PLMs) can capture different levels of concepts in context and hence generate universal language representations. They can benefit from multiple downstream natural language processing (NLP) tasks. Although PTMs have been widely used in most NLP applications, especially for high-resource languages such as English, it is under-represented in Lao NLP research. Previous work on Lao has been hampered by the lack of annotated datasets and the sparsity of language resources. In this work, we construct a text classification dataset to alleviate the resource-scarce situation of the Lao language. In addition, we present the first transformer-based PTMs for Lao with four versions: BERT-Small , BERT-Base , ELECTRA-Small , and ELECTRA-Base . Furthermore, we evaluate them on two downstream tasks: part-of-speech (POS) tagging and text classification. Experiments demonstrate the effectiveness of our Lao models. We release our models and datasets to the community, hoping to facilitate the future development of Lao NLP applications.},
  url   = {https://aclanthology.org/2022.lrec-1.698}
}

@InProceedings{eid-seyffarth-plag:2022:LREC,
  author   = {Eid, Ghattas and Seyffarth, Esther and Plag, Ingo},
  title   = {The Maaloula Aramaic Speech Corpus (MASC): From Printed Material to a Lemmatized and Time-Aligned Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6513--6520},
  abstract = {This paper presents the first electronic speech corpus of Maaloula Aramaic, an endangered Western Neo-Aramaic variety spoken in Syria. This 64,845-word corpus is available in

four formats: (1) transcriptions, (2) lemmatized transcriptions, (3) audio files and time-aligned phonetic transcriptions, and (4) an SQLite database. The transcription files are a digitized and corrected version of authentic transcriptions of tape-recorded narratives coming from a fieldwork trip conducted in the 1980s and published in the early 1990s (Arnold, 1991a, 1991b). They contain no annotation, except for some informative tagging (e.g. to mark loanwords and misspoken words). In the lemmatized version of the files, each word form is followed by its lemma in angled brackets. The time-aligned TextGrid annotations consist of four tiers: the sentence level (Tier 1), the word level (Tiers 2 and 3), and the segment level (Tier 4). These TextGrid files are downloadable together with their audio files (for the original source of the audio data see Arnold, 2003). The SQLite database enables users to access the data on the level of tokens, types, lemmas, sentences, narratives, or speakers. The corpus is now available to the scientific community at https://doi.org/10.5281/zenodo.6496714.},
  url   = {https://aclanthology.org/2022.lrec-1.699}
}

@InProceedings{bond-choo:2022:LREC,
  author   = {Bond, Francis and Choo, Merrick},
  title   = {Sense and Sentiment},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {61--69},
  abstract = {In this paper we examine existing sentiment lexicons and sense-based sentiment-tagged corpora to find out how sense and concept-based semantic relations effect sentiment scores (for polarity and valence). We show that some relations are good predictors of sentiment of related words: antonyms have similar valence and opposite polarity, synonyms similar valence and polarity, as do many derivational relations. We use this knowledge and existing resources to build a sentiment annotated wordnet of English, and show how it can be used to produce sentiment lexicons for other languages using the Open Multilingual Wordnet.},
  url   = {https://aclanthology.org/2022.lrec-1.7}
}

@InProceedings{jacobsen-mohtaj-mller:2022:LREC,
  author   = {Jacobsen, Anik and Mohtaj, Salar and Möller, Sebastian},
  title   = {MuLVE, A Multi-Language Vocabulary Evaluation Data Set},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},

  pages   = {673--679},
  abstract = {Vocabulary learning is vital to foreign language learning. Correct and adequate feedback is essential to successful and satisfying vocabulary training. However, many vocabulary and language evaluation systems perform on simple rules and do not account for real-life user learning data. This work introduces Multi-Language Vocabulary Evaluation Data Set (MuLVE), a data set consisting of vocabulary cards and real-life user answers, labeled indicating whether the user answer is correct or incorrect. The data source is user learning data from the Phase6 vocabulary trainer. The data set contains vocabulary questions in German and English, Spanish, and French as target language and is available in four different variations regarding pre-processing and deduplication. We experiment to fine-tune pre-trained BERT language models on the downstream task of vocabulary evaluation with the proposed MuLVE data set. The results provide outstanding results of > 95.5 accuracy and F2-score. The data set is available on the European Language Grid.},
  url   = {https://aclanthology.org/2022.lrec-1.70}
}

@InProceedings{le-EtAl:2022:LREC,
  author   = {Le, Khang and Nguyen, Hien and Le Thanh, Tung and Nguyen, Minh},
  title   = {VIMQA: A Vietnamese Dataset for Advanced Reasoning and Explainable Multi-hop Question Answering},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6521--6529},
  abstract = {Vietnamese is the native language of over 98 million people in the world. However, existing Vietnamese Question Answering (QA) datasets do not explore the model's ability to perform advanced reasoning and provide evidence to explain the answer. We introduce VIMQA, a new Vietnamese dataset with over 10,000 Wikipedia-based multi-hop question-answer pairs. The dataset is human-generated and has four main features: (1) The questions require advanced reasoning over multiple paragraphs. (2) Sentence-level supporting facts are provided, enabling the QA model to reason and explain the answer. (3) The dataset offers various types of reasoning to test the model's ability to reason and extract relevant proof. (4) The dataset is in Vietnamese, a low-resource language. We also conduct experiments on our dataset using state-of-the-art Multilingual single-hop and multi-hop QA methods. The results suggest that our dataset is challenging for existing methods, and there is room for improvement in Vietnamese QA systems. In addition, we propose a general process for data creation and publish a framework for creating multilingual multi-hop QA datasets. The dataset and framework are publicly available to encourage further research in Vietnamese QA systems.},
  url   = {https://aclanthology.org/2022.lrec-1.700}

}

@InProceedings{dunn-nijhof:2022:LREC,
  author   = {Dunn, Jonathan and Nijhof, Wikke},
  title   = {Language Identification for Austronesian Languages},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6530--6539},
  abstract = {This paper provides language identification models
for low- and under-resourced languages in the Pacific region with a
focus on previously unavailable Austronesian languages. Accurate
language identification is an important part of developing language
resources. The approach taken in this paper combines 29 Austronesian
languages with 171 non-Austronesian languages to create an
evaluation set drawn from eight data sources. After evaluating six
approaches to language identification, we find that a classifier
based on skip-gram embeddings reaches a significantly higher
performance than alternate methods. We then systematically increase
the number of non-Austronesian languages in the model up to a total
of 800 languages to evaluate whether an increased language inventory
leads to less precise predictions for the Austronesian languages of
interest. This evaluation finds that there is only a minimal impact
on accuracy caused by increasing the inventory of non-Austronesian
languages. Further experiments adapt these language identification
models for code-switching detection, achieving high accuracy across
all 29 languages.},
  url   = {https://aclanthology.org/2022.lrec-1.701}
}

@InProceedings{chanda:2022:LREC,
  author   = {Chandía, Andrés},
  title   = {A Mapudüngun FST Morphological Analyser and its Web
Interface},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6540--6547},
  abstract = {This paper describes the development and evaluation
of a FST-based analyser-generator for Mapudüngun language, which is
publicly available through a web interface. As far as we know, it is
the first system of this kind for Mapudüngun. Following the Mapuche
grammar by Smeets, we have developed a machine including the
morphological and phonological aspects of Mapudüngun. Through this
computational approach we have produced a finite state morphological
analyser-generator capable of classifying and appropriately tagging
all the components (roots and suffixes) interacting in a Mapuche
word-form. A double evaluation has been carried out showing a good

level of reliability. In order to face the lack of standardization of the language, additional components (an enhanced analyser, a spelling unifier and a root guesser) have been integrated in the tool. The generated corpora, the lexicons and the FST grammars are available for further development and comparison results.},
  url   = {https://aclanthology.org/2022.lrec-1.702}
}

@InProceedings{cruz-cheng:2022:LREC,
  author   = {Cruz, Jan Christian Blaise and Cheng, Charibeth},
  title   = {Improving Large-scale Language Models and Resources for Filipino},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6548--6555},
  abstract = {In this paper, we improve on existing language resources for the low-resource Filipino language in two ways. First, we outline the construction of the TLUnified dataset, a large-scale pretraining corpus that serves as an improvement over smaller existing pretraining datasets for the language in terms of scale and topic variety. Second, we pretrain new Transformer language models following the RoBERTa pretraining technique to supplant existing models trained with small corpora. Our new RoBERTa models show significant improvements over existing Filipino models in three benchmark datasets with an average gain of 4.47\% test accuracy across three classification tasks with varying difficulty.},
  url   = {https://aclanthology.org/2022.lrec-1.703}
}

@InProceedings{mahadevan-EtAl:2022:LREC,
  author   = {Mahadevan, Shankar and Ponnusamy, Rahul and Kumaresan, Prasanna Kumar and Chandran, Prabakaran and Priyadharshini, Ruba and S, Sangeetha and Chakravarthi, Bharathi Raja},
  title   = {Thirumurai: A Large Dataset of Tamil Shaivite Poems and Classification of Tamil Pann},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6556--6562},
  abstract = {Thirumurai, also known as Panniru Thirumurai, is a collection of Tamil Shaivite poems dating back to the Hindu revival period between the 6th and the 10th century. These poems are par excellence, in both literary and musical terms. They have been composed based on the ancient, now non-existent Tamil Pann system and can be set to music. We present a large dataset containing all the Thirumurai poems and also attempt to classify the Pann and

author of each poem using transformer based architectures. Our work is the first of its kind in dealing with ancient Tamil text datasets, which are severely under-resourced. We explore several Deep Learning-based techniques for solving this challenge effectively and provide essential insights into the problem and how to address it.},
  url   = {https://aclanthology.org/2022.lrec-1.704}
}

@InProceedings{narzary-EtAl:2022:LREC,
  author   = {Narzary, Sanjib and Brahma, Maharaj and Narzary, Mwnthai and Muchahary, Gwmsrang and Singh, Pranav Kumar and Senapati, Apurbalal and Nandi, Sukumar and Som, Bidisha},
  title   = {Generating Monolingual Dataset for Low Resource Language Bodo from old books using Google Keep},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6563--6570},
  abstract = {Bodo is a scheduled Indian language spoken largely by the Bodo community of Assam and other northeastern Indian states. Due to a lack of resources, it is difficult for young languages to communicate more effectively with the rest of the world. This leads to a lack of research in low-resource languages. The creation of a dataset is a tedious and costly process, particularly for languages with no participatory research. This is more visible for languages that are young and have recently adopted standard writing scripts. In this paper, we present a methodology using Google Keep for OCR to generate a monolingual Bodo corpus from different books. In this work, a Bodo text corpus of 192,327 tokens and 32,268 unique tokens is generated using free, accessible, and daily-usable applications. Moreover, some essential characteristics of the Bodo language are discussed that are neglected by Natural Language Progressing (NLP) researchers.},
  url   = {https://aclanthology.org/2022.lrec-1.705}
}

@InProceedings{pathak-nandi-sarmah:2022:LREC,
  author   = {Pathak, Dhrubajyoti and Nandi, Sukumar and Sarmah, Priyankoo},
  title   = {AsNER - Annotated Dataset and Baseline for Assamese Named Entity recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6571--6577},
  abstract = {We present the AsNER, a named entity annotation dataset for low resource Assamese language with a baseline Assamese

NER model. The dataset contains about 99k tokens comprised of text from the speech of the Prime Minister of India and Assamese play. It also contains person names, location names and addresses. The proposed NER dataset is likely to be a significant resource for deep neural based Assamese language processing. We benchmark the dataset by training NER models and evaluating using state-of-the-art architectures for supervised named entity recognition (NER) such as Fasttext, BERT, XLM-R, FLAIR, MuRIL etc. We implement several baseline approaches with state-of-the-art sequence tagging Bi-LSTM-CRF architecture. The highest F1-score among all baselines achieves an accuracy of 80.69\% when using MuRIL as a word embedding method. The annotated dataset and the top performing model are made publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.706}
}

@InProceedings{gaim-yang-park:2022:LREC,
  author   = {Gaim, Fitsum and Yang, Wonsuk and Park, Jong C.},
  title   = {GeezSwitch: Language Identification in Typologically Related Low-resourced East African Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6578--6584},
  abstract = {Language identification is one of the fundamental tasks in natural language processing that is a prerequisite to data processing and numerous applications. Low-resourced languages with similar typologies are generally confused with each other in real-world applications such as machine translation, affecting the user’s experience. In this work, we present a language identification dataset for five typologically and phylogenetically related low-resourced East African languages that use the Ge’ez script as a writing system; namely Amharic, Blin, Ge’ez, Tigre, and Tigrinya. The dataset is built automatically from selected data sources, but we also performed a manual evaluation to assess its quality. Our approach to constructing the dataset is cost-effective and applicable to other low-resource languages. We integrated the dataset into an existing language-identification tool and also fine-tuned several Transformer based language models, achieving very strong results in all cases. While the task of language identification is easy for the informed person, such datasets can make a difference in real-world deployments and also serve as part of a benchmark for language understanding in the target languages. The data and models are made available at https://github.com/fgaim/geezswitch.},
  url   = {https://aclanthology.org/2022.lrec-1.707}
}

@InProceedings{platanou-pavlopoulos-papaioannou:2022:LREC,
  author   = {Platanou, Paraskevi and Pavlopoulos, John and Papaioannou, Georgios},

  title   = {Handwritten Paleographic Greek Text Recognition: A Century-Based Approach},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6585--6589},
  abstract = {Today classicists are provided with a great number of digital tools which, in turn, offer possibilities for further study and new research goals. In this paper we explore the idea that old Greek handwriting can be machine-readable and consequently, researchers can study the target material fast and efficiently. Previous studies have shown that Handwritten Text Recognition (HTR) models are capable of attaining high accuracy rates. However, achieving high accuracy HTR results for Greek manuscripts is still considered to be a major challenge. The overall aim of this paper is to assess HTR for old Greek manuscripts. To address this statement, we study and use digitized images of the Oxford University Bodleian Library Greek manuscripts. By manually transcribing 77 images, we created and present here a new dataset for Handwritten Paleographic Greek Text Recognition. The dataset instances were organized by establishing as a leading factor the century to which the manuscript and hence the image belongs. Experimenting then with an HTR model we show that the error rate depends on the century of the image.},
  url   = {https://aclanthology.org/2022.lrec-1.708}
}

@InProceedings{chida-murakami-pituxcoosuvarn:2022:LREC,
  author   = {Chida, Hiroki and Murakami, Yohei and Pituxcoosuvarn, Mondheera},
  title   = {Quality Control for Crowdsourced Bilingual Dictionary in Low-Resource Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6590--6596},
  abstract = {In conventional bilingual dictionary creation by using crowdsourcing, the main method is to ask multiple workers to translate the same words or sentences and take a majority vote. However, when this method is applied to the creation of bilingual dictionaries for low-resource languages with few speakers, many low-quality workers are expected to participate in the majority voting, which makes it difficult to maintain the quality of the evaluation by the majority voting. Therefore, we apply an effective aggregation method using a hyper question, which is a set of single questions, for quality control. Furthermore, to select high-quality workers, we design a task-allocation method based on the reliability of workers which is evaluated by their work results.},
  url   = {https://aclanthology.org/2022.lrec-1.709}

}

@InProceedings{zilio-EtAl:2022:LREC,
  author   = {Zilio, Leonardo and Saadany, Hadeel and Sharma, Prashant and Kanojia, Diptesh and Orăsan, Constantin},
  title   = {PLOD: An Abbreviation Detection Dataset for Scientific Documents},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {680--688},
  abstract = {The detection and extraction of abbreviations from unstructured texts can help to improve the performance of Natural Language Processing tasks, such as machine translation and information retrieval. However, in terms of publicly available datasets, there is not enough data for training deep-neural-networks-based models to the point of generalising well over data. This paper presents PLOD, a large-scale dataset for abbreviation detection and extraction that contains 160k+ segments automatically annotated with abbreviations and their long forms. We performed manual validation over a set of instances and a complete automatic validation for this dataset. We then used it to generate several baseline models for detecting abbreviations and long forms. The best models achieved an F1-score of 0.92 for abbreviations and 0.89 for detecting their corresponding long forms. We release this dataset along with our code and all the models publicly at https://github.com/surrey-nlp/PLOD-AbbreviationDetection},
  url   = {https://aclanthology.org/2022.lrec-1.71}
}

@InProceedings{oliver-EtAl:2022:LREC,
  author   = {Oliver, Bruce and Forbes, Clarissa and Yang, Changbing and Samir, Farhan and Coates, Edith and Nicolai, Garrett and Silfverberg, Miikka},
  title   = {An Inflectional Database for Gitksan},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6597--6606},
  abstract = {This paper presents a new inflectional resource for Gitksan, a low-resource Indigenous language of Canada. We use Gitksan data in interlinear glossed format, stemming from language documentation efforts, to build a database of partial inflection tables. We then enrich this morphological resource by filling in blank slots in the partial inflection tables using neural transformer reinflection models. We extend the training data for our transformer reinflection models using two data augmentation techniques: data hallucination and back-translation. Experimental

results demonstrate substantial improvements from data augmentation, with data hallucination delivering particularly impressive gains. We also release reinflection models for Gitksan.},
  url   = {https://aclanthology.org/2022.lrec-1.710}
}

@InProceedings{lee-EtAl:2022:LREC4,
  author   = {Lee, Jackson and Chen, Litong and Lam, Charles and Lau, Chaak Ming and Tsui, Tsz-Him},
  title   = {PyCantonese: Cantonese Linguistics and NLP in Python},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6607--6611},
  abstract = {This paper introduces PyCantonese, an open-source Python library for Cantonese linguistics and natural language processing. After the library design, implementation, corpus data format, and key datasets included are introduced, the paper provides an overview of the currently implemented functionality: stop words, handling Jyutping romanization, word segmentation, part-of-speech tagging, and parsing Cantonese text.},
  url   = {https://aclanthology.org/2022.lrec-1.711}
}

@InProceedings{ababu-woldeyohannis:2022:LREC,
  author   = {Ababu, Teshome Mulugeta and Woldeyohannis, Michael Melese},
  title   = {Afaan Oromo Hate Speech Detection and Classification on Social Media},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6612--6619},
  abstract = {Hate and offensive speech on social media is targeted to attack an individual or group of community based on protected characteristics such as gender, ethnicity, and religion. Hate and offensive speech on social media is a global problem that suffers the community especially, for an under-resourced language like Afaan Oromo language. One of the most widely spoken Cushitic language families is Afaan Oromo. Our objective is to develop and test a model used to detect and classify Afaan Oromo hate speech on social media. We developed numerous models that were used to detect and classify Afaan Oromo hate speech on social media by using different machine learning algorithms (classical, ensemble, and deep learning) with the combination of different feature extraction techniques such as BOW, TF-IDF, word2vec, and Keras Embedding layers. To perform the task, we required Afaan Oromo datasets, but the datasets were

unavailable. By concentrating on four thematic areas of hate speech, such as gender, religion, race, and offensive speech, we were able to collect a total of 12,812 posts and comments from Facebook. BiLSTM with pre-trained word2vec feature extraction is an outperformed algorithm that achieves better accuracy of 0.84 and 0.88 for eight classes and two classes, respectively.},
  url   = {https://aclanthology.org/2022.lrec-1.712}
}

@InProceedings{sasano:2022:LREC,
  author   = {Sasano, Ryohei},
  title   = {Cross-lingual Linking of Automatically Constructed Frames and FrameNet},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6620--6625},
  abstract = {A semantic frame is a conceptual structure describing an event, relation, or object along with its participants. Several semantic frame resources have been manually elaborated, and there has been much interest in the possibility of applying semantic frames designed for a particular language to other languages, which has led to the development of cross-lingual frame knowledge. However, manually developing such cross-lingual lexical resources is labor-intensive. To support the development of such resources, this paper presents an attempt at automatic cross-lingual linking of automatically constructed frames and manually crafted frames. Specifically, we link automatically constructed example-based Japanese frames to English FrameNet by using cross-lingual word embeddings and a two-stage model that first extracts candidate FrameNet frames for each Japanese frame by taking only the frame-evoking words into account, then finds the best alignment of frames by also taking frame elements into account. Experiments using frame-annotated sentences in Japanese FrameNet indicate that our approach will facilitate the manual development of cross-lingual frame resources.},
  url   = {https://aclanthology.org/2022.lrec-1.713}
}

@InProceedings{barbu-barbumititelu-mititelu:2022:LREC,
  author   = {Barbu, Ana-Maria and Barbu Mititelu, Verginica and Mititelu, Cătălin},
  title   = {Aligning the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6626--6634},

  abstract = {We present here the efforts of aligning two language resources for Romanian: the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs: for each occurrence of those verbs in the treebank that were included as entries in the lexicon, a set of valence frames is automatically assigned, then manually validated by two linguists and, when necessary, corrected. Validating a valence frame also means semantically disambiguating the verb in the respective context. The validation is done by two linguists, on complementary datasets. However, a subset of verbs were validated by both annotators and Cohen's κ is 0.87 for this subset. The alignment we have made also serves as a method of enhancing the quality of the two resources, as in the process we identify morpho-syntactic annotation mistakes, incomplete valence frames or missing ones. Information from each resource complements the information from the other, thus their value increases. The treebank and the lexicon are freely available, while the links discovered between them are also made available on GitHub.},
  url   = {https://aclanthology.org/2022.lrec-1.714}
}

@InProceedings{lopes-EtAl:2022:LREC,
  author   = {Lopes, Lucelene and Duran, Magali and Fernandes, Paulo and Pardo, Thiago},
  title   = {PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6635--6643},
  abstract = {This paper presents PortiLexicon-UD, a large and freely available lexicon for Portuguese delivering morphosyntactic information according to the Universal Dependencies model. This lexical resource includes part of speech tags, lemmas, and morphological information for words, with 1,221,218 entries (considering word duplication due to different combination of PoS tag, lemma, and morphological features). We report the lexicon creation process, its computational data structure, and its evaluation over an annotated corpus, showing that it has a high language coverage and good quality data.},
  url   = {https://aclanthology.org/2022.lrec-1.715}
}

@InProceedings{gezmu-nrnberger-bati:2022:LREC,
  author   = {Gezmu, Andargachew Mekonnen and Nürnberger, Andreas and Bati, Tesfaye Bayu},
  title   = {Extended Parallel Corpus for Amharic-English Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {6644--6653},
abstract = {This paper describes the acquisition, preprocessing, segmentation, and alignment of an Amharic-English parallel corpus. It will be helpful for machine translation of a low-resource language, Amharic. We freely released the corpus for research purposes. Furthermore, we developed baseline statistical and neural machine translation systems; we trained statistical and neural machine translation models using the corpus. In the experiments, we also used a large monolingual corpus for the language model of statistical machine translation and back-translation of neural machine translation. In the automatic evaluation, neural machine translation models outperform statistical machine translation models by approximately six to seven Bilingual Evaluation Understudy (BLEU) points. Besides, among the neural machine translation models, the subword models outperform the word-based models by three to four BLEU points. Moreover, two other relevant automatic evaluation metrics, Translation Edit Rate on Character Level and Better Evaluation as Ranking, reflect corresponding differences among the trained models.},
url = {https://aclanthology.org/2022.lrec-1.716}
}

@InProceedings{dione-EtAl:2022:LREC,
author = {Dione, Cheikh M. Bamba and LO, Alla and Nguer, Elhadji Mamadou and ba, sileye},
title = {Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof<->French},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {6654--6661},
abstract = {In this paper, we propose two neural machine translation (NMT) systems (French-to-Wolof and Wolof-to-French) based on sequence-to-sequence with attention and Transformer architectures. We trained our models on the parallel French-Wolof corpus (Nguer et al., 2020) of about 83k sentence pairs. Because of the low-resource setting, we experimented with advanced methods for handling data sparsity, including subword segmentation, backtranslation and the copied corpus method. We evaluate the models using BLEU score and find that the transformer outperforms the classic sequence-to-sequence model in all settings, in addition to being less sensitive to noise. In general, the best scores are achieved when training the models on subword-level based units. For such models, using backtranslation proves to be slightly beneficial in low-resource Wolof to high-resource French language translation for the transformer-based models. A slight improvement can also be observed when injecting copied monolingual text in the target language. Moreover, combining the copied method data with backtranslation leads to a slight improvement of the translation

quality.},
  url   = {https://aclanthology.org/2022.lrec-1.717}
}

@InProceedings{demirsahin-EtAl:2022:LREC,
  author   = {Demirsahin, Isin and Johny, Cibu and Gutkin, Alexander and Roark, Brian},
  title   = {Criteria for Useful Automatic Romanization in South Asian Languages},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6662--6673},
  abstract = {This paper presents a number of possible criteria for systems that transliterate South Asian languages from their native scripts into the Latin script, a process known as romanization. These criteria are related to either fidelity to human linguistic behavior (pronunciation transparency, naturalness and conventionality) or processing utility for people (ease of input) as well as under-the-hood in systems (invertibility and stability across languages and scripts). When addressing these differing criteria several linguistic considerations, such as modeling of prominent phonological processes and their relation to orthography, need to be taken into account. We discuss these key linguistic details in the context of Brahmic scripts and languages that use them, such as Hindi and Malayalam. We then present the core features of several romanization algorithms, implemented in a finite state transducer (FST) formalism, that address differing criteria. Implementations of these algorithms have been released as part of the Nisaba finite-state script processing library.},
  url   = {https://aclanthology.org/2022.lrec-1.718}
}

@InProceedings{dai-kamps-sharoff:2022:LREC,
  author   = {Dai, Yuqian and Kamps, Marc de and Sharoff, Serge},
  title   = {BERTology for Machine Translation: What BERT Knows about Linguistic Difficulties for Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6674--6690},
  abstract = {Pre-trained transformer-based models, such as BERT, have shown excellent performance in most natural language processing benchmark tests, but we still lack a good understanding of the linguistic knowledge of BERT in Neural Machine Translation (NMT). Our work uses syntactic probes and Quality Estimation (QE) models to analyze the performance of BERT's syntactic dependencies and their

impact on machine translation quality, exploring what kind of syntactic dependencies are difficult for NMT engines based on BERT. While our probing experiments confirm that pre-trained BERT “knows” about syntactic dependencies, its ability to recognize them often decreases after fine-tuning for NMT tasks. We also detect a relationship between syntactic dependencies in three languages and the quality of their translations, which shows which specific syntactic dependencies are likely to be a significant cause of low-quality translations.},
  url   = {https://aclanthology.org/2022.lrec-1.719}
}

@InProceedings{adewumi-EtAl:2022:LREC,
  author   = {Adewumi, Tosin and Vadoodi, Roshanak and Tripathy, Aparajita and Nikolaido, Konstantina and Liwicki, Foteini and Liwicki, Marcus},
  title   = {Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {689--696},
  abstract = {We present a fairly large, Potential Idiomatic Expression (PIE) dataset for Natural Language Processing (NLP) in English. The challenges with NLP systems with regards to tasks such as Machine Translation (MT), word sense disambiguation (WSD) and information retrieval make it imperative to have a labelled idioms dataset with classes such as it is in this work. To the best of the authors' knowledge, this is the first idioms corpus with classes of idioms beyond the literal and the general idioms classification. In particular, the following classes are labelled in the dataset: metaphor, simile, euphemism, parallelism, personification, oxymoron, paradox, hyperbole, irony and literal. We obtain an overall inter-annotator agreement (IAA) score, between two independent annotators, of 88.89\%. Many past efforts have been limited in the corpus size and classes of samples but this dataset contains over 20,100 samples with almost 1,200 cases of idioms (with their meanings) from 10 classes (or senses). The corpus may also be extended by researchers to meet specific needs. The corpus has part of speech (PoS) tagging from the NLTK library. Classification experiments performed on the corpus to obtain a baseline and comparison among three common models, including the BERT model, give good results. We also make publicly available the corpus and the relevant codes for working with it for NLP tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.72}
}

@InProceedings{jia-EtAl:2022:LREC2,
  author   = {Jia, Ye and Tadmor Ramanovich, Michelle and Wang, Quan and Zen, Heiga},
  title   = {CVSS Corpus and Massively Multilingual Speech-to-

Speech Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6691--6703},
  abstract = {We introduce CVSS, a massively multilingual-to-English speech-to-speech translation (S2ST) corpus, covering sentence-level parallel S2ST pairs from 21 languages into English. CVSS is derived from the Common Voice speech corpus and the CoVoST 2 speech-to-text translation (ST) corpus, by synthesizing the translation text from CoVoST 2 into speech using state-of-the-art TTS systems. Two versions of translation speech in English are provided: 1) CVSS-C: All the translation speech is in a single high-quality canonical voice; 2) CVSS-T: The translation speech is in voices transferred from the corresponding source speech. In addition, CVSS provides normalized translation text which matches the pronunciation in the translation speech. On each version of CVSS, we built baseline multilingual direct S2ST models and cascade S2ST models, verifying the effectiveness of the corpus. To build strong cascade S2ST baselines, we trained an ST model on CoVoST 2, which outperforms the previous state-of-the-art trained on the corpus without extra data by 5.8 BLEU. Nevertheless, the performance of the direct S2ST models approaches the strong cascade baselines when trained from scratch, and with only 0.1 or 0.7 BLEU difference on ASR transcribed translation when initialized from matching ST models.},
  url   = {https://aclanthology.org/2022.lrec-1.720}
}

@InProceedings{morishita-EtAl:2022:LREC,
  author   = {Morishita, Makoto and Chousa, Katsuki and Suzuki, Jun and Nagata, Masaaki},
  title   = {JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6704--6710},
  abstract = {Most current machine translation models are mainly trained with parallel corpora, and their translation accuracy largely depends on the quality and quantity of the corpora. Although there are billions of parallel sentences for a few language pairs, effectively dealing with most language pairs is difficult due to a lack of publicly available parallel corpora. This paper creates a large parallel corpus for English-Japanese, a language pair for which only limited resources are available, compared to such resource-rich languages as English-German. It introduces a new web-based English-Japanese parallel corpus named JParaCrawl v3.0. Our

new corpus contains more than 21 million unique parallel sentence pairs, which is more than twice as many as the previous JParaCrawl v2.0 corpus. Through experiments, we empirically show how our new corpus boosts the accuracy of machine translation models on various domains. The JParaCrawl v3.0 corpus will eventually be publicly available online for research purposes.},
  url   = {https://aclanthology.org/2022.lrec-1.721}
}

@InProceedings{kim-EtAl:2022:LREC4,
  author   = {Kim, Hwichan and Moon, Sangwhan and Okazaki, Naoaki and Komachi, Mamoru},
  title   = {Learning How to Translate North Korean through South Korean},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6711--6718},
  abstract = {South and North Korea both use the Korean language. However, Korean NLP research has focused on South Korean only, and existing NLP systems of the Korean language, such as neural machine translation (NMT) models, cannot properly handle North Korean inputs. Training a model using North Korean data is the most straightforward approach to solving this problem, but there is insufficient data to train NMT models. In this study, we create data for North Korean NMT models using a comparable corpus. First, we manually create evaluation data for automatic alignment and machine translation, and then, investigate automatic alignment methods suitable for North Korean. Finally, we show that a model trained by North Korean bilingual data without human annotation significantly boosts North Korean translation accuracy compared to existing South Korean models in zero-shot settings.},
  url   = {https://aclanthology.org/2022.lrec-1.722}
}

@InProceedings{zhu-EtAl:2022:LREC,
  author   = {Zhu, Wenhao and Huang, Shujian and Pu, Tong and Huang, Pingxuan and zhang, xu and Yu, Jian and Chen, Wei and Wang, Yanfeng and CHEN, Jiajun},
  title   = {FGraDA: A Dataset and Benchmark for Fine-Grained Domain Adaptation in Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6719--6727},
  abstract = {Previous research for adapting a general neural machine translation (NMT) model into a specific domain usually neglects the diversity in translation within the same domain, which

is a core problem for domain adaptation in real-world scenarios. One representative of such challenging scenarios is to deploy a translation system for a conference with a specific topic, e.g., global warming or coronavirus, where there are usually extremely less resources due to the limited schedule. To motivate wider investigation in such a scenario, we present a real-world fine-grained domain adaptation task in machine translation (FGraDA). The FGraDA dataset consists of Chinese-English translation task for four sub-domains of information technology: autonomous vehicles, AI education, real-time networks, and smart phone. Each sub-domain is equipped with a development set and test set for evaluation purposes. To be closer to reality, FGraDA does not employ any in-domain bilingual training data but provides bilingual dictionaries and wiki knowledge base, which can be easier obtained within a short time. We benchmark the fine-grained domain adaptation task and present in-depth analyses showing that there are still challenging problems to further improve the performance with heterogeneous resources.},
  url   = {https://aclanthology.org/2022.lrec-1.723}
}

@InProceedings{nehrdich:2022:LREC,
  author   = {Nehrdich, Sebastian},
  title   = {SansTib, a Sanskrit - Tibetan Parallel Corpus and Bilingual Sentence Embedding Model},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6728--6734},
  abstract = {This paper presents the development of SansTib, a Sanskrit - Classical Tibetan parallel corpus automatically aligned on sentence-level, and a bilingual sentence embedding model. The corpus has a size of about 317,289 sentence pairs and 14,420,771 tokens and thereby is a considerable improvement over previous resources for these two languages. The data is incorporated into the BuddhaNexus database to make it accessible to a larger audience. It also presents a gold evaluation dataset and assesses the quality of the automatic alignment.},
  url   = {https://aclanthology.org/2022.lrec-1.724}
}

@InProceedings{li-EtAl:2022:LREC3,
  author   = {Li, Yihang and Shimizu, Shuichiro and Gu, Weiqi and Chu, Chenhui and Kurohashi, Sadao},
  title   = {VISA: An Ambiguous Subtitles Dataset for Visual Scene-aware Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

publisher = {European Language Resources Association},
pages = {6735--6743},
abstract = {Existing multimodal machine translation (MMT) datasets consist of images and video captions or general subtitles which rarely contain linguistic ambiguity, making visual information not so effective to generate appropriate translations. We introduce VISA, a new dataset that consists of 40k Japanese-English parallel sentence pairs and corresponding video clips with the following key features: (1) the parallel sentences are subtitles from movies and TV episodes; (2) the source subtitles are ambiguous, which means they have multiple possible translations with different meanings; (3) we divide the dataset into Polysemy and Omission according to the cause of ambiguity. We show that VISA is challenging for the latest MMT system, and we hope that the dataset can facilitate MMT research.},
url = {https://aclanthology.org/2022.lrec-1.725}
}

@InProceedings{tani-EtAl:2022:LREC,
author = {Tani, Kazuki and Yuasa, Ryoya and Takikawa, Kazuki and Tamura, Akihiro and Kajiwara, Tomoyuki and Ninomiya, Takashi and Kato, Tsuneo},
title = {A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation},
booktitle = {Proceedings of the Language Resources and Evaluation Conference},
month = {June},
year = {2022},
address = {Marseille, France},
publisher = {European Language Resources Association},
pages = {6744--6752},
abstract = {This paper presents a new benchmark test dataset for multi-level complexity-controllable machine translation (MLCC-MT), which is MT controlling the complexity of the output at more than two levels. In previous research, MLCC-MT models have been evaluated on a test dataset automatically constructed from the Newsela corpus, which is a document-level comparable corpus with document-level complexity. The existing test dataset has the following three problems: (i) A source language sentence and its target language sentence are not necessarily an exact translation pair because they are automatically detected. (ii) A target language sentence and its simplified target language sentence are not necessarily exactly parallel because they are automatically aligned. (iii) A sentence-level complexity is not necessarily appropriate because it is transferred from an article-level complexity attached to the Newsela corpus. Therefore, we create a benchmark test dataset for Japanese-to-English MLCC-MT from the Newsela corpus by introducing an automatic filtering of data with inappropriate sentence-level complexity, manual check for parallel target language sentences with different complexity levels, and manual translation. Moreover, we implement two MLCC-NMT frameworks with a Transformer architecture and report their performance on our test dataset as baselines for future research. Our test dataset and codes are released.},
url = {https://aclanthology.org/2022.lrec-1.726}

}

@InProceedings{lankford-EtAl:2022:LREC,
  author   = {Lankford, Séamus and Afli, Haithem and Ní Loinsigh, Órla and Way, Andy},
  title   = {gaHealth: An English–Irish Bilingual Corpus of Health Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6753--6758},
  abstract = {Machine Translation is a mature technology for many high-resource language pairs. However in the context of low-resource languages, there is a paucity of parallel data datasets available for developing translation models. Furthermore, the development of datasets for low-resource languages often focuses on simply creating the largest possible dataset for generic translation. The benefits and development of smaller in-domain datasets can easily be overlooked. To assess the merits of using in-domain data, a dataset for the specific domain of health was developed for the low-resource English to Irish language pair. Our study outlines the process used in developing the corpus and empirically demonstrates the benefits of using an in-domain dataset for the health domain. In the context of translating health-related data, models developed using the gaHealth corpus demonstrated a maximum BLEU score improvement of 22.2 points (40\%) when compared with top performing models from the LoResMT2021 Shared Task. Furthermore, we define linguistic guidelines for developing gaHealth, the first bilingual corpus of health data for the Irish language, which we hope will be of use to other creators of low-resource data sets. gaHealth is now freely available online and is ready to be explored for further research.},
  url   = {https://aclanthology.org/2022.lrec-1.727}
}

@InProceedings{knowles-littell:2022:LREC,
  author   = {Knowles, Rebecca and Littell, Patrick},
  title   = {Translation Memories as Baselines for Low-Resource Machine Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6759--6767},
  abstract = {Low-resource machine translation research often requires building baselines to benchmark estimates of progress in translation quality. Neural and statistical phrase-based systems are often used with out-of-the-box settings to build these initial baselines before analyzing more sophisticated approaches, implicitly comparing the first machine translation system to the absence of any

translation assistance. We argue that this approach overlooks a basic resource: if you have parallel text, you have a translation memory. In this work, we show that using available text as a translation memory baseline against which to compare machine translation systems is simple, effective, and can shed light on additional translation challenges.},
  url   = {https://aclanthology.org/2022.lrec-1.728}
}

@InProceedings{wang-EtAl:2022:LREC3,
  author   = {Wang, Zhen and Shan, Xu and Zhang, Xiangxie and Yang, Jie},
  title   = {N24News: A New Dataset for Multimodal News Classification},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6768--6775},
  abstract = {Current news datasets merely focus on text features on the news and rarely leverage the feature of images, excluding numerous essential features for news classification. In this paper, we propose a new dataset, N24News, which is generated from New York Times with 24 categories and contains both text and image information in each news. We use a multitask multimodal method and the experimental results show multimodal news classification performs better than text-only news classification. Depending on the length of the text, the classification accuracy can be increased by up to 8.11\%. Our research reveals the relationship between the performance of a multimodal classifier and its sub-classifiers, and also the possible improvements when applying multimodal in news classification. N24News is shown to have great potential to prompt the multimodal news studies.},
  url   = {https://aclanthology.org/2022.lrec-1.729}
}

@InProceedings{bexte-EtAl:2022:LREC,
  author   = {Bexte, Marie and Laarmann-Quante, Ronja and Horbach, Andrea and Zesch, Torsten},
  title   = {LeSpell - A Multi-Lingual Benchmark Corpus of Spelling Errors to Develop Spellchecking Methods for Learner Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {697--706},
  abstract = {Spellchecking text written by language learners is especially challenging because errors made by learners differ both quantitatively and qualitatively from errors made by already

proficient learners. We introduce LeSpell, a multi-lingual (English, German, Italian, and Czech) evaluation data set of spelling mistakes in context that we compiled from seven underlying learner corpora. Our experiments show that existing spellcheckers do not work well with learner data. Thus, we introduce a highly customizable spellchecking component for the DKPro architecture, which improves performance in many settings.},
  url   = {https://aclanthology.org/2022.lrec-1.73}
}

@InProceedings{wang-figueiredo-specia:2022:LREC,
  author   = {Wang, Josiah and Figueiredo, Josiel and Specia, Lucia},
  title   = {MultiSubs: A Large-scale Multimodal and Multilingual Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6776--6785},
  abstract = {This paper introduces a large-scale multimodal and multilingual dataset that aims to facilitate research on grounding words to images in their contextual usage in language. The dataset consists of images selected to unambiguously illustrate concepts expressed in sentences from movie subtitles. The dataset is a valuable resource as (i) the images are aligned to text fragments rather than whole sentences; (ii) multiple images are possible for a text fragment and a sentence; (iii) the sentences are free-form and real-world like; (iv) the parallel texts are multilingual. We also set up a fill-in-the-blank game for humans to evaluate the quality of the automatic image selection process of our dataset. Finally, we propose a fill-in-the-blank task to demonstrate the utility of the dataset, and present some baseline prediction models. The dataset will benefit research on visual grounding of words especially in the context of free-form sentences, and can be obtained from https://doi.org/10.5281/zenodo.5034604 under a Creative Commons licence.},
  url   = {https://aclanthology.org/2022.lrec-1.730}
}

@InProceedings{dai-EtAl:2022:LREC,
  author   = {Dai, Wenliang and Cahyawijaya, Samuel and Yu, Tiezheng and Barezi, Elham J. and Xu, Peng and YIU, Cheuk Tung and Frieske, Rita and Lovenia, Holy and Winata, Genta and Chen, Qifeng and Ma, Xiaojuan and Shi, Bertram and Fung, Pascale},
  title   = {CI-AVSR: A Cantonese Audio-Visual Speech Datasetfor In-car Command Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {6786--6793},
  abstract = {With the rise of deep learning and intelligent vehicles, the smart assistant has become an essential in-car component to facilitate driving and provide extra functionalities. In-car smart assistants should be able to process general as well as car-related commands and perform corresponding actions, which eases driving and improves safety. However, there is a data scarcity issue for low resource languages, hindering the development of research and applications. In this paper, we introduce a new dataset, Cantonese In-car Audio-Visual Speech Recognition (CI-AVSR), for in-car command recognition in the Cantonese language with both video and audio data. It consists of 4,984 samples (8.3 hours) of 200 in-car commands recorded by 30 native Cantonese speakers. Furthermore, we augment our dataset using common in-car background noises to simulate real environments, producing a dataset 10 times larger than the collected one. We provide detailed statistics of both the clean and the augmented versions of our dataset. Moreover, we implement two multimodal baselines to demonstrate the validity of CI-AVSR. Experiment results show that leveraging the visual signal improves the overall performance of the model. Although our best model can achieve a considerable quality on the clean test set, the speech recognition quality on the noisy data is still inferior and remains an extremely challenging task for real in-car speech recognition systems. The dataset and code will be released at https://github.com/HLTCHKUST/CI-AVSR.},
  url   = {https://aclanthology.org/2022.lrec-1.731}
}

@InProceedings{hojo-EtAl:2022:LREC,
  author   = {Hojo, Nobukatsu and Kobashikawa, Satoshi and Mizuno, Saki and Masumura, Ryo},
  title   = {Multimodal Negotiation Corpus with Various Subjective Assessments for Social-Psychological Outcome Prediction from Non-Verbal Cues},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6794--6801},
  abstract = {This study investigates social-psychological negotiation-outcome prediction (SPNOP), a novel task for estimating various subjective evaluation scores of negotiation, such as satisfaction and trust, from negotiation dialogue data. To investigate SPNOP, a corpus with various psychological measurements is beneficial because the interaction process of negotiation relates to many aspects of psychology. However, current negotiation corpora only include information related to objective outcomes or a single aspect of psychology. In addition, most use the ``laboratory setting'' that uses non-skilled negotiators and over simplified negotiation scenarios. There is a concern that such a gap with actual negotiation will intrinsically affect the behavior and

psychology of negotiators in the corpus, which can degrade the performance of models trained from the corpus in real situations. Therefore, we created a negotiation corpus with three features; 1) was assessed with various psychological measurements, 2) used skilled negotiators, and 3) used scenarios of context-rich negotiation. We recorded video and audio of negotiations in Japanese to investigate SPNOP in the context of social signal processing. Experimental results indicate that social-psychological outcomes can be effectively estimated from multimodal information.},
  url   = {https://aclanthology.org/2022.lrec-1.732}
}

@InProceedings{xu-EtAl:2022:LREC,
  author   = {Xu, Shuo and Jia, Yuxiang and Niu, Changyong and Zan, Hongying},
  title   = {MMDAG: Multimodal Directed Acyclic Graph Network for Emotion Recognition in Conversation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6802--6807},
  abstract = {Emotion recognition in conversation is important for an empathetic dialogue system to understand the user’s emotion and then generate appropriate emotional responses. However, most previous researches focus on modeling conversational contexts primarily based on the textual modality or simply utilizing multimodal information through feature concatenation. In order to exploit multimodal information and contextual information more effectively, we propose a multimodal directed acyclic graph (MMDAG) network by injecting information flows inside modality and across modalities into the DAG architecture. Experiments on IEMOCAP and MELD show that our model outperforms other state-of-the-art models. Comparative studies validate the effectiveness of the proposed modality fusion method.},
  url   = {https://aclanthology.org/2022.lrec-1.733}
}

@InProceedings{jang-EtAl:2022:LREC,
  author   = {Jang, Jin Yea and Park, Han-Mu and Shin, Saim and Shin, Suna and Yoon, Byungcheon and Gweon, Gahgene},
  title   = {Automatic Gloss-level Data Augmentation for Sign Language Translation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6808--6813},
  abstract = {Securing sufficient data to enable automatic sign language translation modeling is challenging. The data insufficiency

issue exists in both video and text modalities; however, fewer studies have been performed on text data augmentation compared to video data. In this study, we present three methods of augmenting sign language text modality data, comprising 3,052 Gloss-level Korean Sign Language (GKSL) and Word-level Korean Language (WKL) sentence pairs. Using each of the three methods, the following number of sentence pairs were created: blank replacement 10,654, sentence paraphrasing 1,494, and synonym replacement 899. Translation experiment results using the augmented data showed that when translating from GKSL to WKL and from WKL to GKSL, Bi-Lingual Evaluation Understudy (BLEU) scores improved by 0.204 and 0.170 respectively, compared to when only the original data was used. The three contributions of this study are as follows. First, we demonstrated that three different augmentation techniques used in existing Natural Language Processing (NLP) can be applied to sign language. Second, we propose an automatic data augmentation method which generates quality data by utilizing the Korean sign language gloss dictionary. Lastly, we publish the Gloss-level Korean Sign Language 13k dataset (GKSL13k), which has verified data quality through expert reviews.},
  url   = {https://aclanthology.org/2022.lrec-1.734}
}

@InProceedings{tanaka-EtAl:2022:LREC,
  author   = {Tanaka, Kento and Nishimura, Taichi and Nanjo, Hiroaki and Shirai, Keisuke and Kameko, Hirotaka and Dantsuji, Masatake},
  title   = {Image Description Dataset for Language Learners},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6814--6821},
  abstract = {We focus on image description and a corresponding assessment system for language learners. To achieve automatic assessment of image description, we construct a novel dataset, the Language Learner Image Description (LLID) dataset, which consists of images, their descriptions, and assessment annotations. Then, we propose a novel task of automatic error correction for image description, and we develop a baseline model that encodes multimodal information from a learner sentence with an image and accurately decodes a corrected sentence. Our experimental results show that the developed model can revise errors that cannot be revised without an image.},
  url   = {https://aclanthology.org/2022.lrec-1.735}
}

@InProceedings{cardoso-cohn:2022:LREC,
  author   = {Cardoso, Bruno and Cohn, Neil},
  title   = {The Multimodal Annotation Software Tool (MAST)},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},

  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6822--6828},
  abstract = {Multimodal combinations of writing and pictures have become ubiquitous in contemporary society, and scholars have increasingly been turning to analyzing these media. Here we present a resource for annotating these complex documents: the Multimodal Annotation Software Tool (MAST). MAST is an application that allows users to analyze visual and multimodal documents by selecting and annotating visual regions, and to establish relations between annotations that create dependencies and/or constituent structures. By means of schema publications, MAST allows annotation theories to be citable, while evolving and being shared. Documents can be annotated using multiple schemas simultaneously, offering more comprehensive perspectives. As a distributed, client-server system MAST allows for collaborative annotations across teams of users, and features team management and resource access functionalities, facilitating the potential for implementing open science practices. Altogether, we aim for MAST to provide a powerful and innovative annotation tool with application across numerous fields engaging with multimodal media.},
  url   = {https://aclanthology.org/2022.lrec-1.736}
}

@InProceedings{schwiebert-EtAl:2022:LREC,
  author   = {Schwiebert, Gerald and Weber, Cornelius and Qu, Leyuan and Siqueira, Henrique and Wermter, Stefan},
  title   = {A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6829--6836},
  abstract = {Large datasets as required for deep learning of lip reading do not exist in many languages. In this paper we present the dataset GLips (German Lips) consisting of 250,000 publicly available videos of the faces of speakers of the Hessian Parliament, which was processed for word-level lip reading using an automatic pipeline. The format is similar to that of the English language LRW (Lip Reading in the Wild) dataset, with each video encoding one word of interest in a context of 1.16 seconds duration, which yields compatibility for studying transfer learning between both datasets. By training a deep neural network, we investigate whether lip reading has language-independent features, so that datasets of different languages can be used to improve lip reading models. We demonstrate learning from scratch and show that transfer learning from LRW to GLips and vice versa improves learning speed and performance, in particular for the validation set.},
  url   = {https://aclanthology.org/2022.lrec-1.737}

}

@InProceedings{garg-EtAl:2022:LREC2,
  author   = {Garg, Muskan and Wazarkar, Seema and Singh, Muskaan and Bojar, Ondřej},
  title   = {Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6837--6847},
  abstract = {With the development of multimodal systems and natural language generation techniques, the resurgence of multimodal datasets has attracted significant research interests, which aims to provide new information to enrich the representation of textual data. However, there remains a lack of a comprehensive survey for this task. To this end, we take the first step and present a thorough review of this research field. This paper provides an overview of a publicly available dataset with different modalities according to the applications. Furthermore, we discuss the new frontier and give our thoughts. We hope this survey of multimodal datasets can provide the community with quick access and a general picture of the multimodal dataset for specific Natural Language Processing (NLP) applications and motivates future researches. In this context, we release the collection of all multimodal datasets easily accessible here: https://github.com/drmuskangarg/Multimodal-datasets},
  url   = {https://aclanthology.org/2022.lrec-1.738}
}

@InProceedings{carlsson-EtAl:2022:LREC,
  author   = {Carlsson, Fredrik and Eisen, Philipp and Rekathati, Faton and Sahlgren, Magnus},
  title   = {Cross-lingual and Multilingual CLIP},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6848--6854},
  abstract = {The long-standing endeavor of relating the textual and the visual domain recently underwent a pivotal breakthrough, as OpenAI released CLIP. This model distinguishes how well an English text corresponds with a given image with unprecedented accuracy. Trained via a contrastive learning objective over a huge dataset of 400M of images and captions, it is a work that is not easily replicated, especially for low resource languages. Capitalizing on the modularization of the CLIP architecture, we propose to use cross-lingual teacher learning to re-train the textual encoder for various non-English languages. Our method requires no image data and

relies entirely on machine translation which removes the need for data in the target language. We find that our method can efficiently train a new textual encoder with relatively low computational cost, whilst still outperforming previous baselines on multilingual image-text retrieval.},
  url   = {https://aclanthology.org/2022.lrec-1.739}
}

@InProceedings{seiffe-EtAl:2022:LREC,
  author   = {Seiffe, Laura and Kallel, Fares and Möller, Sebastian and Naderi, Babak and Roller, Roland},
  title   = {Subjective Text Complexity Assessment for German},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {707--714},
  abstract = {For different reasons, text can be difficult to read and understand for many people, especially if the text's language is too complex. In order to provide suitable text for the target audience, it is necessary to measure its complexity. In this paper we describe subjective experiments to assess the readability of German text. We compile a new corpus of sentences provided by a German IT service provider. The sentences are annotated with the subjective complexity ratings by two groups of participants, namely experts and non-experts for that text domain. We then extract an extensive set of linguistically motivated features that are supposedly interacting with complexity perception. We show that a linear regression model with a subset of these features can be a very good predictor of text complexity.},
  url   = {https://aclanthology.org/2022.lrec-1.74}
}

@InProceedings{khan-shifath-islam:2022:LREC,
  author   = {Khan, Mohammad Faiyaz and Shifath, S.M. Sadiq-Ur-Rahman and Islam, Md Saiful},
  title   = {BAN-Cap: A Multi-Purpose English-Bangla Image Descriptions Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6855--6865},
  abstract = {As computers have become efficient at understanding visual information and transforming it into a written representation, research interest in tasks like automatic image captioning has seen a significant leap over the last few years. While most of the research attention is given to the English language in a monolingual setting, resource-constrained languages like Bangla remain out of focus, predominantly due to a lack of

standard datasets. Addressing this issue, we present a new dataset BAN-Cap following the widely used Flickr8k dataset, where we collect Bangla captions of the images provided by qualified annotators. Our dataset represents a wider variety of image caption styles annotated by trained people from different backgrounds. We present a quantitative and qualitative analysis of the dataset and the baseline evaluation of the recent models in Bangla image captioning. We investigate the effect of text augmentation and demonstrate that an adaptive attention-based model combined with text augmentation using Contextualized Word Replacement (CWR) outperforms all state-of-the-art models for Bangla image captioning. We also present this dataset's multipurpose nature, especially on machine translation for Bangla-English and English-Bangla. This dataset and all the models will be useful for further research.},
  url   = {https://aclanthology.org/2022.lrec-1.740}
}

@InProceedings{kimura-EtAl:2022:LREC,
  author   = {Kimura, Naoki and Su, Zixiong and Saeki, Takaaki and Rekimoto, Jun},
  title   = {SSR7000: A Synchronized Corpus of Ultrasound Tongue Imaging for End-to-End Silent Speech Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6866--6873},
  abstract = {This article presents SSR7000, a corpus of synchronized ultrasound tongue and lip images designed for end-to-end silent speech recognition (SSR). Although neural end-to-end models are successfully updating the state-of-the-art technology in the field of automatic speech recognition, SSR research based on ultrasound tongue imaging has still not evolved past cascaded DNN-HMM models due to the absence of a large dataset. In this study, we constructed a large dataset, namely SSR7000, to exploit the performance of the end-to-end models. The SSR7000 dataset contains ultrasound tongue and lip images of 7484 utterances by a single speaker. It contains more utterances per person than any other SSR corpus based on ultrasound imaging. We also describe preprocessing techniques to tackle data variances that are inevitable when collecting a large dataset and present benchmark results using an end-to-end model. The SSR7000 corpus is publicly available under the CC BY-NC 4.0 license.},
  url   = {https://aclanthology.org/2022.lrec-1.741}
}

@InProceedings{zhao-EtAl:2022:LREC,
  author   = {Zhao, Yang and Kanayama, Hiroshi and Yoshida, Issei and Muraoka, Masayasu and Aizawa, Akiko},
  title   = {A Simple Yet Effective Corpus Construction Method for Chinese Sentence Compression},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6874--6883},
  abstract = {Deletion-based sentence compression in the English language has made significant progress over the past few decades. However, there is a lack of large-scale and high-quality parallel corpus (i.e., (sentence, compression) pairs) for the Chinese language to train an efficient compression system. To remedy this shortcoming, we present a dependency-tree-based method to construct a Chinese corpus with 151k pairs of sentences and compression based on Chinese language-specific characteristics. Subsequently, we trained both extractive and generative neural compression models using the constructed corpus. The experimental results show that our compression model can generate high-quality compressed sentences on both automatic and human evaluation metrics compared with the baselines. The results of the faithfulness evaluation also indicated that the Chinese compression model trained on our constructed corpus can produce more faithful compressed sentences. Furthermore, a dataset with 1,000 pairs of sentences and ground truth compression was manually created for automatic evaluation, which, we believe, will benefit future research on Chinese sentence compression.},
  url   = {https://aclanthology.org/2022.lrec-1.742}
}

@InProceedings{huang-kajiwara-arase:2022:LREC,
  author   = {Huang, Han and Kajiwara, Tomoyuki and Arase, Yuki},
  title   = {{JADE}: Corpus for Japanese Definition Modelling},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6884--6888},
  abstract = {This study investigated and released the JADE, a corpus for Japanese definition modelling, which is a technique that automatically generates definitions of a given target word and phrase. It is a crucial technique for practical applications that assist language learning and education, as well as for those supporting reading documents in unfamiliar domains. Although corpora for development of definition modelling techniques have been actively created, their languages are mostly limited to English. In this study, a corpus for Japanese, named JADE, was created following the previous study that mines an online encyclopedia. The JADE provides about 630k sets of targets, their definitions, and usage examples as contexts for about 41k unique targets, which is sufficiently large to train neural models. The targets are both words and phrases, and the coverage of domains and topics is diverse. The performance of a pre-trained sequence-to-sequence model and the state-of-the-art definition modelling method was also

benchmarked on JADE for future development of the technique in Japanese. The JADE corpus has been released and available online.},
  url   = {https://aclanthology.org/2022.lrec-1.743}
}

@InProceedings{pu-EtAl:2022:LREC,
  author   = {Pu, Jiashu and Huang, Ziyi and Xi, Yadong and Chen, Guandan and Chen, Weijie and Zhang, Rongsheng},
  title   = {Unraveling the Mystery of Artifacts in Machine Generated Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6889--6898},
  abstract = {As neural Text Generation Models (TGM) have become more and more capable of generating text indistinguishable from human-written ones, the misuse of text generation technologies can have serious ramifications. Although a neural classifier often achieves high detection accuracy, the reason for it is not well studied. Most previous work revolves around studying the impact of model structure and the decoding strategy on ease of detection, but little work has been done to analyze the forms of artifacts left by the TGM. We propose to systematically study the forms and scopes of artifacts by corrupting texts, replacing them with linguistic or statistical features, and applying the interpretable method of Integrated Gradients. Comprehensive experiments show artifacts a) primarily relate to token co-occurrence, b) feature more heavily at the head of vocabulary, c) appear more in content word than stopwords, d) are sometimes detrimental in the form of number of token occurrences, e) are less likely to exist in high-level semantics or syntaxes, f) manifest in low concreteness values for higher-order n-grams.},
  url   = {https://aclanthology.org/2022.lrec-1.744}
}

@InProceedings{chang-EtAl:2022:LREC,
  author   = {Chang, Ernie and Kovtunova, Alisa and Borgwardt, Stefan and Demberg, Vera and Chapman, Kathryn and Yeh, Hui-Syuan},
  title   = {Logic-Guided Message Generation from Raw Real-Time Sensor Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6899--6908},
  abstract = {Natural language generation in real-time settings with raw sensor data is a challenging task. We find that formulating the task as an end-to-end problem leads to two major challenges in

content selection -- the sensor data is both redundant and diverse across environments, thereby making it hard for the encoders to select and reason on the data. We here present a new corpus for a specific domain that instantiates these properties. It includes handover utterances that an assistant for a semi-autonomous drone uses to communicate with humans during the drone flight. The corpus consists of sensor data records and utterances in 8 different environments. As a structured intermediary representation between data records and text, we explore the use of description logic (DL). We also propose a neural generation model that can alert the human pilot of the system state and environment in preparation of the handover of control.},
  url   = {https://aclanthology.org/2022.lrec-1.745}
}

@InProceedings{kumar-EtAl:2022:LREC2,
  author   = {Kumar, Ayush and Jani, Dhyey and Shah, Jay and Thakar, Devanshu and Jain, Varun and Singh, Mayank},
  title   = {The Bull and the Bear: Summarizing Stock Market Discussions},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6909--6913},
  abstract = {Stock market investors debate and heavily discuss stock ideas, investing strategies, news and market movements on social media platforms. The discussions are significantly longer in length and require extensive domain expertise for understanding. In this paper, we curate such discussions and construct a first-of-its-kind of abstractive summarization dataset. Our curated dataset consists of 7888 Reddit posts and manually constructed summaries for 400 posts. We robustly evaluate the summaries and conduct experiments on SOTA summarization tools to showcase their limitations. We plan to make the dataset publicly available. The sample dataset is available here: https://dhyeyjani.github.io/RSMC},
  url   = {https://aclanthology.org/2022.lrec-1.746}
}

@InProceedings{espasa-morin-hamon:2022:LREC,
  author   = {Espasa, Kévin and Morin, Emmanuel and Hamon, Olivier},
  title   = {Combination of Contextualized and Non-Contextualized Layers for Lexical Substitution in French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6914--6921},
  abstract = {Lexical substitution task requires to substitute a

target word by candidates in a given context. Candidates must keep meaning and grammatically of the sentence. The task, introduced in the SemEval 2007, has two objectives. The first objective is to find a list of substitutes for a target word. This list of substitutes can be obtained with lexical resources like WordNet or generated with a pre-trained language model. The second objective is to rank these substitutes using the context of the sentence. Most of the methods use vector space models or more recently embeddings to rank substitutes. Embedding methods use high contextualized representation. This representation can be over contextualized and in this way overlook good substitute candidates which are more similar on non-contextualized layers. SemDis 2014 introduced the lexical substitution task in French. We propose an application of the state-of-the-art method based on BERT in French and a novel method using contextualized and non-contextualized layers to increase the suggestion of words having a lower probability in a given context but that are more semantically similar. Experiments show our method increases the BERT based system on the OOT measure but decreases on the BEST measure in the SemDis 2014 benchmark.},
  url   = {https://aclanthology.org/2022.lrec-1.747}
}

@InProceedings{bastan-EtAl:2022:LREC,
  author   = {Bastan, Mohaddeseh and Shankar, Nishant and Surdeanu, Mihai and Balasubramanian, Niranjan},
  title   = {SuMe: A Dataset Towards Summarizing Biomedical Mechanisms},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6922--6931},
  abstract = {Can language models read biomedical texts and explain the biomedical mechanisms discussed? In this work we introduce a biomedical mechanism summarization task. Biomedical studies often investigate the mechanisms behind how one entity (e.g., a protein or a chemical) affects another in a biological context. The abstracts of these publications often include a focused set of sentences that present relevant supporting statements regarding such relationships, associated experimental evidence, and a concluding sentence that summarizes the mechanism underlying the relationship. We leverage this structure and create a summarization task, where the input is a collection of sentences and the main entities in an abstract, and the output includes the relationship and a sentence that summarizes the mechanism. Using a small amount of manually labeled mechanism sentences, we train a mechanism sentence classifier to filter a large biomedical abstract collection and create a summarization dataset with 22k instances. We also introduce conclusion sentence generation as a pretraining task with 611k instances. We benchmark the performance of large bio-domain language models. We find that while the pretraining task help improves performance, the best model produces acceptable mechanism outputs in only 32\% of the instances,

which shows the task presents significant challenges in biomedical language understanding and summarization.},
  url   = {https://aclanthology.org/2022.lrec-1.748}
}

@InProceedings{chen-lin:2022:LREC,
  author   = {chen, zheng and Lin, Hongyu},
  title   = {CATAMARAN: A Cross-lingual Long Text Abstractive Summarization Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6932--6937},
  abstract = {Cross-lingual summarization, which produces the summary in one language from a given source document in another language, could be extremely helpful for humans to obtain information across the world. However, it is still a little-explored task due to the lack of datasets. Recent studies are primarily based on pseudo-cross-lingual datasets obtained by translation. Such an approach would inevitably lead to the loss of information in the original document and introduce noise into the summary, thus hurting the overall performance. In this paper, we present CATAMARAN, the first high-quality cross-lingual long text abstractive summarization dataset. It contains about 20,000 parallel news articles and corresponding summaries, all written by humans. The average lengths of articles are 1133.65 for English articles and 2035.33 for Chinese articles, and the average lengths of the summaries are 26.59 and 70.05, respectively. We train and evaluate an mBART-based cross-lingual abstractive summarization model using our dataset. The result shows that, compared with mono-lingual systems, the cross-lingual abstractive summarization system could also achieve solid performance.},
  url   = {https://aclanthology.org/2022.lrec-1.749}
}

@InProceedings{frick-schmidt-helmer:2022:LREC,
  author   = {Frick, Elena and Schmidt, Thomas and Helmer, Henrike},
  title   = {Querying Interaction Structure: Approaches to Overlap in Spoken Language Corpora},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {715--722},
  abstract = {In this paper, we address two problems in indexing and querying spoken language corpora with overlapping speaker contributions. First, we look into how token distance and token precedence can be measured when multiple primary data streams are

available and when transcriptions happen to be tokenized, but are not synchronized with the sound at the level of individual tokens. We propose and experiment with a speaker-based search mode that enables any speaker’s transcription tier to be the basic tokenization layer whereby the contributions of other speakers are mapped to this given tier. Secondly, we address two distinct methods of how speaker overlaps can be captured in the TEI-based ISO Standard for Spoken Language Transcriptions (ISO 24624:2016) and how they can be queried by MTAS – an open source Lucene-based search engine for querying text with multilevel annotations. We illustrate the problems, introduce possible solutions and discuss their benefits and drawbacks.},
  url   = {https://aclanthology.org/2022.lrec-1.75}
}

@InProceedings{sosea-EtAl:2022:LREC,
  author   = {Sosea, Tiberiu and Pham, Chau and Tekle, Alexander and Caragea, Cornelia and Li, Junyi Jessy},
  title   = {Emotion analysis and detection during COVID-19},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6938--6947},
  abstract = {Understanding emotions that people express during large-scale crises helps inform policy makers and first responders about the emotional states of the population as well as provide emotional support to those who need such support. We present CovidEmo, a dataset of ~3,000 English tweets labeled with emotions and temporally distributed across 18 months. Our analyses reveal the emotional toll caused by COVID-19, and changes of the social narrative and associated emotions over time. Motivated by the time-sensitive nature of crises and the cost of large-scale annotation efforts, we examine how well large pre-trained language models generalize across domains and timeline in the task of perceived emotion prediction in the context of COVID-19. Our analyses suggest that cross-domain information transfers occur, yet there are still significant gaps. We propose semi-supervised learning as a way to bridge this gap, obtaining significantly better performance using unlabeled data from the target domain.},
  url   = {https://aclanthology.org/2022.lrec-1.750}
}

@InProceedings{hassan-shaar-darwish:2022:LREC,
  author   = {Hassan, Sabit and Shaar, Shaden and Darwish, Kareem},
  title   = {Cross-lingual Emotion Detection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {6948--6958},
  abstract = {Emotion detection can provide us with a window into understanding human behavior. Due to the complex dynamics of human emotions, however, constructing annotated datasets to train automated models can be expensive. Thus, we explore the efficacy of cross-lingual approaches that would use data from a source language to build models for emotion detection in a target language. We compare three approaches, namely: i) using inherently multilingual models; ii) translating training data into the target language; and iii) using an automatically tagged parallel corpus. In our study, we consider English as the source language with Arabic and Spanish as target languages. We study the effectiveness of different classification models such as BERT and SVMs trained with different features. Our BERT-based monolingual models that are trained on target language data surpass state-of-the-art (SOTA) by 4\% and 5\% absolute Jaccard score for Arabic and Spanish respectively. Next, we show that using cross-lingual approaches with English data alone, we can achieve more than 90\% and 80\% relative effectiveness of the Arabic and Spanish BERT models respectively. Lastly, we use LIME to analyze the challenges of training cross-lingual models for different language pairs.},
  url   = {https://aclanthology.org/2022.lrec-1.751}
}

@InProceedings{zhang-liu:2022:LREC,
  author   = {Zhang, Yuanchi and Liu, Yang},
  title   = {DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6959--6966},
  abstract = {Quotation extraction and attribution are challenging tasks, aiming at determining the spans containing quotations and attributing each quotation to the original speaker. Applying this task to news data is highly related to fact-checking, media monitoring and news tracking. Direct quotations are more traceable and informative, and therefore of great significance among different types of quotations. Therefore, this paper introduces DirectQuote, a corpus containing 19,760 paragraphs and 10,279 direct quotations manually annotated from online news media. To the best of our knowledge, this is the largest and most complete corpus that focuses on direct quotations in news texts. We ensure that each speaker in the annotation can be linked to a specific named entity on Wikidata, benefiting various downstream tasks. In addition, for the first time, we propose several sequence labeling models as baseline methods to extract and attribute quotations simultaneously in an end-to-end manner.},
  url   = {https://aclanthology.org/2022.lrec-1.752}
}

@InProceedings{weinzierl-harabagiu:2022:LREC,
  author   = {Weinzierl, Maxwell and Harabagiu, Sanda},
  title   = {VaccineLies: A Natural Language Resource for Learning to Recognize Misinformation about the COVID-19 and HPV Vaccines},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6967--6975},
  abstract = {Billions of COVID-19 vaccines have been administered, but many remain hesitant. Misinformation about the COVID-19 vaccines and other vaccines, propagating on social media, is believed to drive hesitancy towards vaccination. The ability to automatically recognize misinformation targeting vaccines on Twitter depends on the availability of data resources. In this paper we present VaccineLies, a large collection of tweets propagating misinformation about two vaccines: the COVID-19 vaccines and the Human Papillomavirus (HPV) vaccines. Misinformation targets are organized in vaccine-specific taxonomies, which reveal the misinformation themes and concerns. The ontological commitments of the misinformation taxonomies provide an understanding of which misinformation themes and concerns dominate the discourse about the two vaccines covered in VaccineLies. The organization into training, testing and development sets of VaccineLies invites the development of novel supervised methods for detecting misinformation on Twitter and identifying the stance towards it. Furthermore, VaccineLies can be a stepping stone for the development of datasets focusing on misinformation targeting additional vaccines.},
  url   = {https://aclanthology.org/2022.lrec-1.753}
}

@InProceedings{turban-kruschwitz:2022:LREC,
  author   = {Turban, Christoph and Kruschwitz, Udo},
  title   = {Tackling Irony Detection using Ensemble Classifiers},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6976--6984},
  abstract = {Automatic approaches to irony detection have been of interest to the NLP community for a long time, yet, state-of-the-art approaches still fall way short of what one would consider a desirable performance. In part this is due to the inherent difficulty of the problem. However, in recent years ensembles of transformer-based approaches have emerged as a promising direction to push the state of the art forward in a wide range of NLP applications. A different, more recent, development is the automatic augmentation of training data. In this paper we will explore both these directions for the task of irony detection in social media.

Using the common SemEval 2018 Task 3 benchmark collection we demonstrate that transformer models are well suited in ensemble classifiers for the task at hand. In the multi-class classification task we observe statistically significant improvements over strong baselines. For binary classification we achieve performance that is on par with state-of-the-art alternatives. The examined data augmentation strategies showed an effect, but are not decisive for good results.},
  url   = {https://aclanthology.org/2022.lrec-1.754}
}

@InProceedings{ayemar-shirai:2022:LREC,
  author   = {Aye Mar, Aye and Shirai, Kiyoaki},
  title   = {Automatic Construction of an Annotated Corpus with Implicit Aspects},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6985--6991},
  abstract = {Aspect-based sentiment analysis (ABSA) is a task that involves classifying the polarity of aspects of the products or services described in users' reviews. Most previous work on ABSA has focused on explicit aspects, which appear as explicit words or phrases in the sentences of the review. However, users often express their opinions toward the aspects indirectly or implicitly, in which case the specific name of an aspect does not appear in the review. The current datasets used for ABSA are mainly annotated with explicit aspects. This paper proposes a novel method for constructing a corpus that is automatically annotated with implicit aspects. The main idea is that sentences containing explicit and implicit aspects share a similar context. First, labeled sentences with explicit aspects and unlabeled sentences that include implicit aspects are collected. Next, clustering is performed on these sentences so that similar sentences are merged into the same cluster. Finally, the explicit aspects are propagated to the unlabeled sentences in the same cluster, in order to construct a labeled dataset containing implicit aspects. The results of our experiments on mobile phone reviews show that our method of identifying the labels of implicit aspects achieves a maximum accuracy of 82\%.},
  url   = {https://aclanthology.org/2022.lrec-1.755}
}

@InProceedings{ray-EtAl:2022:LREC,
  author   = {Ray, Anupama and Mishra, Shubham and Nunna, Apoorva and Bhattacharyya, Pushpak},
  title   = {A Multimodal Corpus for Emotion Recognition in Sarcasm},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {6992--7003},
  abstract = {While sentiment and emotion analysis have been studied extensively, the relationship between sarcasm and emotion has largely remained unexplored. A sarcastic expression may have a variety of underlying emotions. For example, “I love being ignored” belies sadness, while “my mobile is fabulous with a battery backup of only 15 minutes!” expresses frustration. Detecting the emotion behind a sarcastic expression is non-trivial yet an important task. We undertake the task of detecting the emotion in a sarcastic statement, which to the best of our knowledge, is hitherto unexplored. We start with the recently released multimodal sarcasm detection dataset (MUStARD) pre-annotated with 9 emotions. We identify and correct 343 incorrect emotion labels (out of 690). We double the size of the dataset, label it with emotions along with valence and arousal which are important indicators of emotional intensity. Finally, we label each sarcastic utterance with one of the four sarcasm types-Propositional, Embedded, Likeprefixed and Illocutionary, with the goal of advancing sarcasm detection research. Exhaustive experimentation with multimodal (text, audio, and video) fusion models establishes a benchmark for exact emotion recognition in sarcasm and outperforms the state-of-art sarcasm detection. We release the dataset enriched with various annotations and the code for research purposes: https://github.com/apoorva-nunna/MUStARD_Plus_Plus},
  url   = {https://aclanthology.org/2022.lrec-1.756}
}

@InProceedings{tammewar-EtAl:2022:LREC,
  author   = {Tammewar, Aniruddha and Braun, Franziska and Roccabruna, Gabriel and Bayerl, Sebastian and Riedhammer, Korbinian and Riccardi, Giuseppe},
  title   = {Annotation of Valence Unfolding in Spoken Personal Narratives},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7004--7013},
  abstract = {Personal Narrative (PN) is the recollection of individuals’ life experiences, events, and thoughts along with the associated emotions in the form of a story. Compared to other genres such as social media texts or microblogs, where people write about experienced events or products, the spoken PNs are complex to analyze and understand. They are usually long and unstructured, involving multiple and related events, characters as well as thoughts and emotions associated with events, objects, and persons. In spoken PNs, emotions are conveyed by changing the speech signal characteristics as well as the lexical content of the narrative. In this work, we annotate a corpus of spoken personal narratives, with

the emotion valence using discrete values. The PNs are segmented into speech segments, and the annotators annotate them in the discourse context, with values on a 5-point bipolar scale ranging from -2 to +2 (0 for neutral). In this way, we capture the unfolding of the PNs events and changes in the emotional state of the narrator. We perform an in-depth analysis of the inter-annotator agreement, the relation between the label distribution w.r.t. the stimulus (positive/negative) used for the elicitation of the narrative, and compare the segment-level annotations to a baseline continuous annotation. We find that the neutral score plays an important role in the agreement. We observe that it is easy to differentiate the positive from the negative valence while the confusion with the neutral label is high. Keywords: Personal Narratives, Emotion Annotation, Segment Level Annotation},
  url   = {https://aclanthology.org/2022.lrec-1.757}
}

@InProceedings{nakayama-EtAl:2022:LREC,
  author   = {Nakayama, Yuki and Murakami, Koji and Kumar, Gautam and Bhingardive, Sudha and Hardaway, Ikuko},
  title   = {A Large-Scale Japanese Dataset for Aspect-based Sentiment Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7014--7021},
  abstract = {There has been significant progress in the field of sentiment analysis. However, aspect-based sentiment analysis (ABSA) has not been explored in the Japanese language even though it has a huge scope in many natural language processing applications such as 1) tracking sentiment towards products, movies, politicians etc; 2) improving customer relation models. The main reason behind this is that there is no standard Japanese dataset available for ABSA task. In this paper, we present the first standard Japanese dataset for the hotel reviews domain. The proposed dataset contains 53,192 review sentences with seven aspect categories and two polarity labels. We perform experiments on this dataset using popular ABSA approaches and report error analysis. Our experiments show that contextual models such as BERT works very well for the ABSA task in the Japanese language and also show the need to focus on other NLP tasks for better performance through our error analysis.},
  url   = {https://aclanthology.org/2022.lrec-1.758}
}

@InProceedings{suzuki-EtAl:2022:LREC2,
  author   = {Suzuki, Haruya and Miyauchi, Yuto and Akiyama, Kazuki and Kajiwara, Tomoyuki and Ninomiya, Takashi and Takemura, Noriko and Nakashima, Yuta and Nagahara, Hajime},
  title   = {A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain},
  booktitle   = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7022--7028},
  abstract = {We annotate 35,000 SNS posts with both the writer's subjective sentiment polarity labels and the reader's objective ones to construct a Japanese sentiment analysis dataset. Our dataset includes intensity labels (\textit{none}, \textit{weak}, \textit{medium}, and \textit{strong}) for each of the eight basic emotions by Plutchik (\textit{joy}, \textit{sadness}, \textit{anticipation}, \textit{surprise}, \textit{anger}, \textit{fear}, \textit{disgust}, and \textit{trust}) as well as sentiment polarity labels (\textit{strong positive}, \textit{positive}, \textit{neutral}, \textit{negative}, and \textit{strong negative}). Previous studies on emotion analysis have studied the analysis of basic emotions and sentiment polarity independently. In other words, there are few corpora that are annotated with both basic emotions and sentiment polarity. Our dataset is the first large-scale corpus to annotate both of these emotion labels, and from both the writer's and reader's perspectives. In this paper, we analyze the relationship between basic emotion intensity and sentiment polarity on our dataset and report the results of benchmarking sentiment polarity classification.},
  url   = {https://aclanthology.org/2022.lrec-1.759}
}

@InProceedings{pzik-EtAl:2022:LREC,
  author   = {Pęzik, Piotr and Krawentek, Gosia and Karasińska, Sylwia and Wilk, Paweł and Rybińska, Paulina and Cichosz, Anna and Peljak-Łapińska, Angelika and Deckert, Mikołaj and Adamczyk, Michał},
  title   = {DiaBiz – an Annotated Corpus of Polish Call Center Dialogs},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {723--726},
  abstract = {This paper introduces DiaBiz, a large, annotated, multimodal corpus of Polish telephone conversations conducted in varied business settings, comprising 4036 call centre interactions from nine different domains, i.e. banking, energy services, telecommunications, insurance, medical care, debt collection, tourism, retail and car rental. The corpus was developed to boost the development of third-party speech recognition engines, dialog systems and conversational intelligence tools for Polish. Its current size amounts to nearly 410 hours of recordings and over 3 million words of transcribed speech. We present the structure of the corpus, data collection and transcription procedures, challenges of

punctuating and truecasing speech transcripts, dialog structure annotation and discuss some of the ecological validity considerations involved in the development of such resources.},
  url   = {https://aclanthology.org/2022.lrec-1.76}
}

@InProceedings{qin-EtAl:2022:LREC,
  author   = {Qin, Han and Tian, Yuanhe and Xia, Fei and Song, Yan},
  title   = {Complementary Learning of Aspect Terms for Aspect-based Sentiment Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7029--7039},
  abstract = {Aspect-based sentiment analysis (ABSA) aims to predict the sentiment polarity towards a given aspect term in a sentence on the fine-grained level, which usually requires a good understanding of contextual information, especially appropriately distinguishing of a given aspect and its contexts, to achieve good performance. However, most existing ABSA models pay limited attention to the modeling of the given aspect terms and thus result in inferior results when a sentence contains multiple aspect terms with contradictory sentiment polarities. In this paper, we propose to improve ABSA by complementary learning of aspect terms, which serves as a supportive auxiliary task to enhance ABSA by explicitly recovering the aspect terms from each input sentence so as to better understand aspects and their contexts. Particularly, a discriminator is also introduced to further improve the learning process by appropriately balancing the impact of aspect recovery to sentiment prediction. Experimental results on five widely used English benchmark datasets for ABSA demonstrate the effectiveness of our approach, where state-of-the-art performance is observed on all datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.760}
}

@InProceedings{bose-su:2022:LREC,
  author   = {bose, saugata and Su, Dr. Guoxin},
  title   = {Deep One-Class Hate Speech Detection Model},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7040--7048},
  abstract = {Hate speech detection for social media posts is considered as a binary classification problem in existing approaches, largely neglecting distinct attributes of hate speeches from other sentimental types such as ``aggressive'' and ``racist''.

As these sentimental types constitute a significant major portion of data, the classification performance is compromised. Moreover, those classifiers often do not generalize well across different datasets due to a relatively small number of hate-class samples. In this paper, we adopt a one-class perspective for hate speech detection, where the detection classifier is trained with hate-class samples only. Our model employs a BERT-BiLSTM module for feature extraction and a one-class SVM for classification. A comprehensive evaluation with four benchmarking datasets demonstrates the better performance of our model than existing approaches, as well as the advantage of training our model with a combination of the four datasets.},
  url   = {https://aclanthology.org/2022.lrec-1.761}
}

@InProceedings{barriere-essid-clavel:2022:LREC,
  author   = {Barriere, Valentin and Essid, Slim and Clavel, Chloé},
  title   = {Opinions in Interactions : New Annotations of the SEMAINE Database},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7049--7055},
  abstract = {In this paper, we present the process we used in order to collect new annotations of opinions over the multimodal corpus SEMAINE composed of dyadic interactions. The dataset had already been annotated continuously in two affective dimensions related to the emotions: Valence and Arousal. We annotated the part of SEMAINE called \textit{Solid SAL} composed of 79 interactions between a user and an operator playing the role of a virtual agent designed to engage a person in a sustained, emotionally colored conversation. We aligned the audio at the word level using the available high-quality manual transcriptions. The annotated dataset contains 5627 speech turns for a total of 73,944 words, corresponding to 6 hours 20 minutes of dyadic interactions. Each interaction has been labeled by three annotators at the speech turn level following a three-step process. This method allows us to obtain a precise annotation regarding the opinion of a speaker. We obtain thus a dataset dense in opinions, with more than 48\% of the annotated speech turns containing at least one opinion. We then propose a new baseline for the detection of opinions in interactions improving slightly a state of the art model with RoBERTa embeddings. The obtained results on the database are promising with a F1-score at 0.72.},
  url   = {https://aclanthology.org/2022.lrec-1.762}
}

@InProceedings{shangipourataei-EtAl:2022:LREC,
  author   = {Shangipour ataei, Taha and Darvishi, Kamyar and Javdan, Soroush and Minaei-Bidgoli, Behrouz and Eetemadi, Sauleh},

  title   = {Pars-ABSA: a Manually Annotated Aspect-based Sentiment Analysis Benchmark on Farsi Product Reviews},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7056--7060},
  abstract = {Due to the increased availability of online reviews, sentiment analysis witnessed a thriving interest from researchers. Sentiment analysis is a computational treatment of sentiment used to extract and understand the opinions of authors. While many systems were built to predict the sentiment of a document or a sentence, many others provide the necessary detail on various aspects of the entity (i.e., aspect-based sentiment analysis). Most of the available data resources were tailored to English and the other popular European languages. Although Farsi is a language with more than 110 million speakers, to the best of our knowledge, there is a lack of proper public datasets on aspect-based sentiment analysis for Farsi. This paper provides a manually annotated Farsi dataset, Pars-ABSA, annotated and verified by three native Farsi speakers. The dataset consists of 5,114 positive, 3,061 negative and 1,827 neutral data samples from 5,602 unique reviews. Moreover, as a baseline, this paper reports the performance of some aspect-based sentiment analysis methods focusing on transfer learning on Pars-ABSA.},
  url   = {https://aclanthology.org/2022.lrec-1.763}
}

@InProceedings{-EtAl:2022:LREC,
  author   = {., Mamta and Ekbal, Asif and Bhattacharyya, Pushpak and Saha, Tista and Kumar, Alka and Srivastava, Shikha},
  title   = {HindiMD: A Multi-domain Corpora for Low-resource Sentiment Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7061--7070},
  abstract = {Social media platforms such as Twitter have evolved into a vast information sharing platform, allowing people from a variety of backgrounds and expertise to share their opinions on numerous events such as terrorism, narcotics and many other social issues. People sometimes misuse the power of social media for their agendas, such as illegal trades and negatively influencing others. Because of this, sentiment analysis has won the interest of a lot of researchers to widely analyze public opinion for social media monitoring. Several benchmark datasets for sentiment analysis across a range of domains have been made available, especially for high-resource languages. A few datasets are available for low-resource

Indian languages like Hindi, such as movie reviews and product reviews, which do not address the current need for social media monitoring. In this paper, we address the challenges of sentiment analysis in Hindi and socially relevant domains by introducing a balanced corpus annotated with the sentiment classes, viz. positive, negative and neutral. To show the effective usage of the dataset, we build several deep learning based models and establish them as the baselines for further research in this direction.},
  url   = {https://aclanthology.org/2022.lrec-1.764}
}

@InProceedings{pavlopoulos-xenos-picca:2022:LREC,
  author   = {Pavlopoulos, John and Xenos, Alexandros and Picca, Davide},
  title   = {Sentiment Analysis of Homeric Text: The 1st Book of Iliad},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7071--7077},
  abstract = {Sentiment analysis studies are focused more on online customer reviews or social media, and less on literary studies. The problem is greater for ancient languages, where the linguistic expression of sentiments may diverge from modern linguistic forms. This work presents the outcome of a sentiment annotation task of the first Book of Iliad, an ancient Greek poem. The annotators were provided with verses translated into modern Greek and they annotated the perceived emotions and sentiments verse by verse. By estimating the fraction of annotators that found a verse as belonging to a specific sentiment class, we model the poem’s perceived sentiment as a multi-variate time series. By experimenting with a state of the art deep learning masked language model, pre-trained on modern Greek and fine-tuned to estimate the sentiment of our data, we registered a mean squared error of 0.063. This low error indicates that sentiment estimators built on our dataset can potentially be used as mechanical annotators, hence facilitating the distant reading of Homeric text. Our dataset is released for public use.},
  url   = {https://aclanthology.org/2022.lrec-1.765}
}

@InProceedings{safari-EtAl:2022:LREC,
  author   = {Safari, Pegah and Rasooli, Mohammad Sadegh and Moloodi, Amirsaeid and Nourian, Alireza},
  title   = {The Persian Dependency Treebank Made Universal},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7078--7087},

  abstract = {We describe an automatic method for converting the Persian Dependency Treebank (Rasooli et al., 2013) to Universal Dependencies. This treebank contains 29107 sentences. Our experiments along with manual linguistic analysis show that our data is more compatible with Universal Dependencies than the Uppsala Persian Universal Dependency Treebank (Seraji et al., 2016), larger in size and more diverse in vocabulary. Our data brings in labeled attachment F-score of 85.2 in supervised parsing. Also, our delexicalized Persian-to-English parser transfer experiments show that a parsing model trained on our data is ≈2\% absolutely more accurate than that of Seraji et al. (2016) in terms of labeled attachment score.},
  url   = {https://aclanthology.org/2022.lrec-1.766}
}

@InProceedings{baxi-bhatt:2022:LREC,
  author   = {Baxi, Jatayu and bhatt, brijesh},
  title   = {GujMORPH - A Dataset for Creating Gujarati Morphological Analyzer},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7088--7095},
  abstract = {Computational morphology deals with the processing of a language at the word level. A morphological analyzer is a key linguistic word-level tool that returns all the constituent morphemes and their grammatical categories associated with a particular word form. For the highly inflectional and low resource languages, the creation of computational morphology-related tools is a challenging task due to the unavailability of underlying key resources. In this paper, we discuss the creation of an annotated morphological dataset- GujMORPH for the Gujarati - an indo-aryan language. For the creation of this dataset, we studied language grammar, word formation rules, and suffix attachments in depth. This dataset contains 16,527 unique inflected words along with their morphological segmentation and grammatical feature tagging information. It is a first of its kind dataset for the Gujarati language and can be used to develop morphological analyzer and generator models. The dataset is annotated in the standard Unimorph schema and evaluated on the baseline system. We also describe the tool used to annotate the data in the standard format. The dataset is released publicly along with the library. Using this library, the data can be obtained in a format that can be directly used to train any machine learning model.},
  url   = {https://aclanthology.org/2022.lrec-1.767}
}

@InProceedings{kabiri-karimi-surdeanu:2022:LREC,
  author   = {Kabiri, Roya and Karimi, Simin and Surdeanu, Mihai},
  title   = {Informal Persian Universal Dependency Treebank},

  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7096--7105},
  abstract = {This paper presents the phonological, morphological, and syntactic distinctions between formal and informal Persian, showing that these two variants have fundamental differences that cannot be attributed solely to pronunciation discrepancies. Given that informal Persian exhibits particular characteristics, any computational model trained on formal Persian is unlikely to transfer well to informal Persian, necessitating the creation of dedicated treebanks for this variety. We thus detail the development of the open-source Informal Persian Universal Dependency Treebank, a new treebank annotated within the Universal Dependencies scheme. We then investigate the parsing of informal Persian by training two dependency parsers on existing formal treebanks and evaluating them on out-of-domain data, i.e. the development set of our informal treebank. Our results show that parsers experience a substantial performance drop when we move across the two domains, as they face more unknown tokens and structures and fail to generalize well. Furthermore, the dependency relations whose performance deteriorates the most represent the unique properties of the informal variant. The ultimate goal of this study that demonstrates a broader impact is to provide a stepping-stone to reveal the significance of informal variants of languages, which have been widely overlooked in natural language processing tools across languages.},
  url   = {https://aclanthology.org/2022.lrec-1.768}
}

@InProceedings{zupon-EtAl:2022:LREC,
  author   = {Zupon, Andrew and Carnie, Andrew and Hammond, Michael and Surdeanu, Mihai},
  title   = {Automatic Correction of Syntactic Dependency Annotation Differences},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7106--7112},
  abstract = {Annotation inconsistencies between data sets can cause problems for low-resource NLP, where noisy or inconsistent data cannot be easily replaced. We propose a method for automatically detecting annotation mismatches between dependency parsing corpora, along with three related methods for automatically converting the mismatches. All three methods rely on comparing unseen examples in a new corpus with similar examples in an existing corpus. These three methods include a simple lexical replacement using the most frequent tag of the example in the existing corpus, a GloVe embedding-based replacement that considers related examples,

and a BERT-based replacement that uses contextualized embeddings to provide examples fine-tuned to our data. We evaluate these conversions by retraining two dependency parsers---Stanza and Parsing as Tagging (PaT)---on the converted and unconverted data. We find that applying our conversions yields significantly better performance in many cases. Some differences observed between the two parsers are observed. Stanza has a more complex architecture with a quadratic algorithm, taking longer to train, but it can generalize from less data. The PaT parser has a simpler architecture with a linear algorithm, speeding up training but requiring more training data to reach comparable or better performance.},
  url   = {https://aclanthology.org/2022.lrec-1.769}
}

@InProceedings{daris-EtAl:2022:LREC,
  author   = {Darģis, Roberts and Auziņa, Ilze and Kaija, Inga and Levāne-Petrova, Kristīne and Pokratniece, Kristīne},
  title   = {LaVA – Latvian Language Learner corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {727--731},
  abstract = {This paper presents the Latvian Language Learner Corpus (LaVA) developed at the Institute of Mathematics and Computer Science, University of Latvia. LaVA corpus contains 1015 essays (190k tokens and 790k characters excluding whitespaces) from foreigners studying at Latvian higher education institutions and who are learning Latvian as a foreign language in the first or second semester, reaching the A1 (possibly A2) Latvian language proficiency level. The corpus has morphological and error annotations. Error analysis and the statistics of the LaVA corpus are also provided in the paper. The corpus is publicly available at: http://www.korpuss.lv/id/LaVA.},
  url   = {https://aclanthology.org/2022.lrec-1.77}
}

@InProceedings{sato-yoshinaga-kitsuregawa:2022:LREC,
  author   = {Sato, Fumikazu and Yoshinaga, Naoki and Kitsuregawa, Masaru},
  title   = {Building Large-Scale Japanese Pronunciation-Annotated Corpora for Reading Heteronymous Logograms},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7113--7121},
  abstract = {Although screen readers enable visually impaired people to read written text via speech, the ambiguities in pronunciations of heteronyms cause wrong reading, which has a

serious impact on the text understanding. Especially in Japanese, there are many common heteronyms expressed by logograms (Chinese characters or kanji) that have totally different pronunciations (and meanings). In this study, to improve the accuracy of pronunciation prediction, we construct two large-scale Japanese corpora that annotate kanji characters with their pronunciations. Using existing language resources on i) book titles compiled by the National Diet Library and ii) the books in a Japanese digital library called Aozora Bunko and their Braille translations, we develop two large-scale pronunciation-annotated corpora for training pronunciation prediction models. We first extract sentence-level alignments between the Aozora Bunko text and its pronunciation converted from the Braille data. We then perform dictionary-based pattern matching based on morphological dictionaries to find word-level pronunciation alignments. We have ultimately obtained the Book Title corpus with 336M characters (16.4M book titles) and the Aozora Bunko corpus with 52M characters (1.6M sentences). We analyzed pronunciation distributions for 203 common heteronyms, and trained a BERT-based pronunciation prediction model for 93 heteronyms, which achieved an average accuracy of 0.939.},
  url   = {https://aclanthology.org/2022.lrec-1.770}
}

@InProceedings{cho-EtAl:2022:LREC,
  author   = {Cho, Won Ik and Moon, Sangwhan and Kim, Jongin and Kim, Seokmin and Kim, Nam Soo},
  title   = {StyleKQC: A Style-Variant Paraphrase Corpus for Korean Questions and Commands},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7122--7128},
  abstract = {Paraphrasing is often performed with less concern for controlled style conversion. Especially for questions and commands, style-variant paraphrasing can be crucial in tone and manner, which also matters with industrial applications such as dialog systems. In this paper, we attack this issue with a corpus construction scheme that simultaneously considers the core content and style of directives, namely intent and formality, for the Korean language. Utilizing manually generated natural language queries on six daily topics, we expand the corpus to formal and informal sentences by human rewriting and transferring. We verify the validity and industrial applicability of our approach by checking the adequate classification and inference performance that fit with conventional fine-tuning approaches, at the same time proposing a supervised formality transfer task.},
  url   = {https://aclanthology.org/2022.lrec-1.771}
}

@InProceedings{tian-EtAl:2022:LREC2,
  author   = {Tian, Yuanhe and Qin, Han and Xia, Fei and

Song, Yan},
  title   = {Syntax-driven Approach for Semantic Role Labeling},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7129--7139},
  abstract = {As an important task to analyze the semantic structure of a sentence, semantic role labeling (SRL) aims to locate the semantic role (e.g., agent) of noun phrases with respect to a given predicate and thus plays an important role in downstream tasks such as dialogue systems. To achieve a better performance in SRL, a model is always required to have a good understanding of the context information. Although one can use advanced text encoder (e.g., BERT) to capture the context information, extra resources are also required to further improve the model performance. Considering that there are correlations between the syntactic structure and the semantic structure of the sentence, many previous studies leverage auto-generated syntactic knowledge, especially the dependencies, to enhance the modeling of context information through graph-based architectures, where limited attention is paid to other types of auto-generated knowledge. In this paper, we propose map memories to enhance SRL by encoding different types of auto-generated syntactic knowledge (i.e., POS tags, syntactic constituencies, and word dependencies) obtained from off-the-shelf toolkits. Experimental results on two English benchmark datasets for span-style SRL (i.e., CoNLL-2005 and CoNLL-2012) demonstrate the effectiveness of our approach, which outperforms strong baselines and achieves state-of-the-art results on CoNLL-2005.},
  url   = {https://aclanthology.org/2022.lrec-1.772}
}

@InProceedings{woliski-EtAl:2022:LREC,
  author   = {Woliński, Marcin and Nitoń, Bartłomiej and Kieraś, Witold and Szymanik, Jakub},
  title   = {HerBERT Based Language Model Detects Quantifiers and Their Semantic Properties in Polish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7140--7146},
  abstract = {The paper presents a tool for automatic marking up of quantifying expressions, their semantic features, and scopes. We explore the idea of using a BERT based neural model for the task (in this case HerBERT, a model trained specifically for Polish, is used). The tool is trained on a recent manually annotated Corpus of Polish Quantificational Expressions (Szymanik and Kieraś, 2022). We discuss how it performs against human annotation and present results of automatic annotation of 300 million sub-corpus of National Corpus

of Polish. Our results show that language models can effectively recognise semantic category of quantification as well as identify key semantic properties of quantifiers, like monotonicity. Furthermore, the algorithm we have developed can be used for building semantically annotated quantifier corpora for other languages.},
  url   = {https://aclanthology.org/2022.lrec-1.773}
}

@InProceedings{bao-hauer-kondrak:2022:LREC,
  author   = {Bao, hongchang and Hauer, Bradley and Kondrak, Grzegorz},
  title   = {Lexical Resource Mapping via Translations},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7147--7154},
  abstract = {Aligning lexical resources that associate words with concepts in multiple languages increases the total amount of semantic information that can be leveraged for various NLP tasks. We present a translation-based approach to mapping concepts across diverse resources. Our methods depend only on multilingual lexicalization information. When applied to align WordNet/BabelNet to CLICS and OmegaWiki, our methods achieve state-of-the-art accuracy, without any dependence on other sources of semantic knowledge. Since each word-concept pair corresponds to a unique sense of the word, we also demonstrate that the mapping task can be framed as word sense disambiguation. To facilitate future work, we release a set of high-precision WordNet-CLICS alignments, produced by combining three different mapping methods.},
  url   = {https://aclanthology.org/2022.lrec-1.774}
}

@InProceedings{takahashi-bollegala:2022:LREC,
  author   = {Takahashi, Keigo and Bollegala, Danushka},
  title   = {Unsupervised Attention-based Sentence-Level Meta-Embeddings from Contextualised Language Models},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7155--7163},
  abstract = {A variety of contextualised language models have been proposed in the NLP community, which are trained on diverse corpora to produce numerous Neural Language Models (NLMs). However, different NLMs have reported different levels of performances in downstream NLP applications when used as text representations. We propose a sentence-level meta-embedding learning method that takes independently trained contextualised word embedding models and

learns a sentence embedding that preserves the complementary strengths of the input source NLMs. Our proposed method is unsupervised and is not tied to a particular downstream task, which makes the learnt meta-embeddings in principle applicable to different tasks that require sentence representations. Specifically, we first project the token-level embeddings obtained by the individual NLMs and learn attention weights that indicate the contributions of source embeddings towards their token-level meta-embeddings. Next, we apply mean and max pooling to produce sentence-level meta-embeddings from token-level meta-embeddings. Experimental results on semantic textual similarity benchmarks show that our proposed unsupervised sentence-level meta-embedding method outperforms previously proposed sentence-level meta-embedding methods as well as a supervised baseline.},
  url   = {https://aclanthology.org/2022.lrec-1.775}
}

@InProceedings{khanal-traskowsky-caragea:2022:LREC,
  author   = {Khanal, Sarthak and Traskowsky, Maria and
Caragea, Doina},
  title   = {Identification of Fine-Grained Location Mentions in
Crisis Tweets},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7164--7173},
  abstract = {Identification of fine-grained location mentions in
crisis tweets is central in transforming situational awareness
information extracted from social media into actionable information.
Most prior works have focused on identifying generic locations,
without considering their specific types. To facilitate progress on
the fine-grained location identification task, we assemble two tweet
crisis datasets and manually annotate them with specific location
types. The first dataset contains tweets from a mixed set of crisis
events, while the second dataset contains tweets from the global
COVID-19 pandemic. We investigate the performance of state-of-the-
art deep learning models for sequence tagging on these datasets, in
both in-domain and cross-domain settings.},
  url   = {https://aclanthology.org/2022.lrec-1.776}
}

@InProceedings{vargas-EtAl:2022:LREC2,
  author   = {Vargas, Francielle and Carvalho, Isabelle and
Rodrigues de Góes, Fabiana and Pardo, Thiago and Benevenuto,
Fabrício},
  title   = {HateBR: A Large Expert Annotated Corpus of Brazilian
Instagram Comments for Offensive Language and Hate Speech
Detection},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7174--7183},
  abstract = {Due to the severity of the social media offensive and hateful comments in Brazil, and the lack of research in Portuguese, this paper provides the first large-scale expert annotated corpus of Brazilian Instagram comments for hate speech and offensive language detection. The HateBR corpus was collected from the comment section of Brazilian politicians' accounts on Instagram and manually annotated by specialists, reaching a high inter-annotator agreement. The corpus consists of 7,000 documents annotated according to three different layers: a binary classification (offensive versus non-offensive comments), offensiveness-level classification (highly, moderately, and slightly offensive), and nine hate speech groups (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology for the dictatorship, antisemitism, and fatphobia). We also implemented baseline experiments for offensive language and hate speech detection and compared them with a literature baseline. Results show that the baseline experiments on our corpus outperform the current state-of-the-art for the Portuguese language.},
  url   = {https://aclanthology.org/2022.lrec-1.777}
}

@InProceedings{ji-EtAl:2022:LREC,
  author   = {Ji, Shaoxiong and Zhang, Tianlin and Ansari, Luna and Fu, Jie and Tiwari, Prayag and Cambria, Erik},
  title   = {MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7184--7190},
  abstract = {Mental health is a critical issue in modern society, and mental disorders could sometimes turn to suicidal ideation without adequate treatment. Early detection of mental disorders and suicidal ideation from social content provides a potential way for effective social intervention. Recent advances in pretrained contextualized language representations have promoted the development of several domainspecific pretrained models and facilitated several downstream applications. However, there are no existing pretrained language models for mental healthcare. This paper trains and release two pretrained masked language models, i.e., MentalBERT and MentalRoBERTa, to benefit machine learning for the mental healthcare research community. Besides, we evaluate our trained domain-specific models and several variants of pretrained language models on several mental disorder detection benchmarks and demonstrate that language representations pretrained in the target domain improve the performance of mental health detection tasks.},
  url   = {https://aclanthology.org/2022.lrec-1.778}

}

@InProceedings{liao:2022:LREC,
  author   = {Liao, Yu Yun},
  title   = {Leveraging Hashtag Networks for Multimodal Popularity Prediction of Instagram Posts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7191--7198},
  abstract = {With the increasing commercial and social importance of Instagram in recent years, more researchers begin to take multimodal approaches to predict popular content on Instagram. However, existing popularity prediction approaches often reduce hashtags to simple features such as hashtag length or number of hashtags in a post, ignoring the structural and textual information that entangles between hashtags. In this paper, we propose a multimodal framework using post captions, image, hashtag network, and topic model to predict popular influencer posts in Taiwan. Specifically, the hashtag network is constructed as a homogenous graph using the co-occurrence relationship between hashtags, and we extract its structural information with GraphSAGE and semantic information with BERTopic. Finally, the prediction process is defined as a binary classification task (popular/unpopular) using neural networks. Our results show that the proposed framework incorporating hashtag network outperforms all baselines and unimodal models, while information captured from the hashtag network and topic model appears to be complementary.},
  url   = {https://aclanthology.org/2022.lrec-1.779}
}

@InProceedings{heafield-EtAl:2022:LREC,
  author   = {Heafield, Kenneth and Farrow, Elaine and van der Linde, Jelmer and Ramírez-Sánchez, Gema and Wiggins, Dion},
  title   = {The EuroPat Corpus: A Parallel Corpus of European Patent Data},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {732--740},
  abstract = {We present the EuroPat corpus of patent-specific parallel data for 6 official European languages paired with English: German, Spanish, French, Croatian, Norwegian, and Polish. The filtered parallel corpora range in size from 51 million sentences (Spanish-English) to 154k sentences (Croatian-English), with the unfiltered (raw) corpora being up to 2 times larger. Access to clean, high quality, parallel data in technical domains such as science, engineering, and medicine is needed for training neural

machine translation systems for tasks like online dispute resolution and eProcurement. Our evaluation found that the addition of EuroPat data to a generic baseline improved the performance of machine translation systems on in-domain test data in German, Spanish, French, and Polish; and in translating patent data from Croatian to English. The corpus has been released under Creative Commons Zero, and is expected to be widely useful for training high-quality machine translation systems, and particularly for those targeting technical documents such as patents and contracts.},
  url   = {https://aclanthology.org/2022.lrec-1.78}
}

@InProceedings{jiang-EtAl:2022:LREC2,
  author   = {Jiang, Hang and Hua, Yining and Beeferman, Doug and Roy, Deb},
  title   = {Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7199--7208},
  abstract = {Social media data such as Twitter messages ("tweets") pose a particular challenge to NLP systems because of their short, noisy, and colloquial nature. Tasks such as Named Entity Recognition (NER) and syntactic parsing require highly domain-matched training data for good performance. To date, there is no complete training corpus for both NER and syntactic analysis (e.g., part of speech tagging, dependency parsing) of tweets. While there are some publicly available annotated NLP datasets of tweets, they are only designed for individual tasks. In this study, we aim to create Tweebank-NER, an English NER corpus based on Tweebank V2 (TB2), train state-of-the-art (SOTA) Tweet NLP models on TB2, and release an NLP pipeline called Twitter-Stanza. We annotate named entities in TB2 using Amazon Mechanical Turk and measure the quality of our annotations. We train the Stanza pipeline on TB2 and compare with alternative NLP frameworks (e.g., FLAIR, spaCy) and transformer-based models. The Stanza tokenizer and lemmatizer achieve SOTA performance on TB2, while the Stanza NER tagger, part-of-speech (POS) tagger, and dependency parser achieve competitive performance against non-transformer models. The transformer-based models establish a strong baseline in Tweebank-NER and achieve the new SOTA performance in POS tagging and dependency parsing on TB2. We release the dataset and make both the Stanza pipeline and BERTweet-based models available "off-the-shelf" for use in future Tweet NLP research. Our source code, data, and pre-trained models are available at: \url{https://github.com/social-machines/TweebankNLP}.},
  url   = {https://aclanthology.org/2022.lrec-1.780}
}

@InProceedings{andy-EtAl:2022:LREC,

  author   = {Andy, Anietie and Kriz, Reno and Guntuku, Sharath Chandra and Wijaya, Derry Tanti and Callison-Burch, Chris},
  title   = {Did that happen? Predicting Social Media Posts that are Indicative of what happened in a scene: A case study of a TV show},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7209--7214},
  abstract = {While popular Television (TV) shows are airing, some users interested in these shows publish social media posts about the show. Analyzing social media posts related to a TV show can be beneficial for gaining insights about what happened during scenes of the show. This is a challenging task partly because a significant number of social media posts associated with a TV show or event may not clearly describe what happened during the event. In this work, we propose a method to predict social media posts (associated with scenes of a TV show) that are indicative of what transpired during the scenes of the show. We evaluate our method on social media (Twitter) posts associated with an episode of a popular TV show, Game of Thrones. We show that for each of the identified scenes, with high AUC's, our method can predict posts that are indicative of what happened in a scene from those that are not-indicative. Based on Twitters policy, we will make the Tweeter ID's of the Twitter posts used for this work publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.781}
}

@InProceedings{kodali-EtAl:2022:LREC,
  author   = {Kodali, Prashant and Bhatnagar, Akshala and Ahuja, Naman and Shrivastava, Manish and Kumaraguru, Ponnurangam},
  title   = {HashSet - A Dataset For Hashtag Segmentation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7215--7219},
  abstract = {Hashtag segmentation is the task of breaking a hashtag into its constituent tokens. Hashtags often encode the essence of user-generated posts, along with information like topic and sentiment, which are useful in downstream tasks. Hashtags prioritize brevity and are written in unique ways - transliterating and mixing languages, spelling variations, creative named entities. Benchmark datasets used for the hashtag segmentation task - STAN, BOUN - are small and extracted from a single set of tweets. However, datasets should reflect the variations in writing styles of hashtags and account for domain and language specificity, failing which the results will misrepresent model performance. We argue that model

performance should be assessed on a wider variety of hashtags, and datasets should be carefully curated. To this end, we propose HashSet, a dataset comprising of: a) 1.9k manually annotated dataset; b) 3.3M loosely supervised dataset. HashSet dataset is sampled from a different set of tweets when compared to existing datasets and provides an alternate distribution of hashtags to build and validate hashtag segmentation models. We analyze the performance of SOTA models for Hashtag Segmentation, and show that the proposed dataset provides an alternate set of hashtags to train and assess models.},
  url   = {https://aclanthology.org/2022.lrec-1.782}
}

@InProceedings{tran-phung-ngo:2022:LREC,
  author   = {Tran, Oanh and Phung, Anh Cong and Ngo, Bach Xuan},
  title   = {Using Convolution Neural Network with BERT for Stance Detection in Vietnamese},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7220--7225},
  abstract = {Stance detection is the task of automatically eliciting stance information towards a specific claim made by a primary author. While most studies have been done for high-resource languages, this work is dedicated to a low-resource language, namely Vietnamese. In this paper, we propose an architecture using transformers to detect stances in Vietnamese claims. This architecture exploits BERT to extract contextual word embeddings instead of using traditional word2vec models. Then, these embeddings are fed into CNN networks to extract local features to train the stance detection model. We performed extensive comparison experiments to show the effectiveness of the proposed method on a public dataset1 Experimental results show that this proposed model outperforms the previous methods by a large margin. It yielded an accuracy score of 75.57\% averaged on four labels. This sets a new SOTA result for future research on this interesting problem in Vietnamese.},
  url   = {https://aclanthology.org/2022.lrec-1.783}
}

@InProceedings{murayama-EtAl:2022:LREC,
  author   = {Murayama, Taichi and Hisada, Shohei and Uehara, Makoto and Wakamiya, Shoko and ARAMAKI, Eiji},
  title   = {Annotation-Scheme Reconstruction for "Fake News" and Japanese Fake News Dataset},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},

  publisher   = {European Language Resources Association},
  pages   = {7226--7234},
  abstract = {Fake news provokes many societal problems; therefore, there has been extensive research on fake news detection tasks to counter it. Many fake news datasets were constructed as resources to facilitate this task. Contemporary research focuses almost exclusively on the factuality aspect of the news. However, this aspect alone is insufficient to explain "fake news," which is a complex phenomenon that involves a wide range of issues. To fully understand the nature of each instance of fake news, it is important to observe it from various perspectives, such as the intention of the false news disseminator, the harmfulness of the news to our society, and the target of the news. We propose a novel annotation scheme with fine-grained labeling based on detailed investigations of existing fake news datasets to capture these various aspects of fake news. Using the annotation scheme, we construct and publish the first Japanese fake news dataset. The annotation scheme is expected to provide an in-depth understanding of fake news. We plan to build datasets for both Japanese and other languages using our scheme. Our Japanese dataset is published at https://hkefka385.github.io/dataset/fakenews-japanese/.},
  url   = {https://aclanthology.org/2022.lrec-1.784}
}

@InProceedings{prez-EtAl:2022:LREC,
  author   = {Pérez, Juan Manuel and Furman, Damián Ariel and Alonso Alemany, Laura and Luque, Franco M.},
  title   = {RoBERTuito: a pre-trained language model for social media text in Spanish},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7235--7243},
  abstract = {Since BERT appeared, Transformer language models and transfer learning have become state-of-the-art for natural language processing tasks. Recently, some works geared towards pre-training specially-crafted models for particular domains, such as scientific papers, medical documents, user-generated texts, among others. These domain-specific models have been shown to improve performance significantly in most tasks; however, for languages other than English, such models are not widely available. In this work, we present RoBERTuito, a pre-trained language model for user-generated text in Spanish, trained on over 500 million tweets. Experiments on a benchmark of tasks involving user-generated text showed that RoBERTuito outperformed other pre-trained language models in Spanish. In addition to this, our model has some cross-lingual abilities, achieving top results for English-Spanish tasks of the Linguistic Code-Switching Evaluation benchmark (LinCE) and also competitive performance against monolingual models in English Twitter tasks. To facilitate further research, we make RoBERTuito publicly available at the HuggingFace model hub together with the

dataset used to pre-train it.},
  url   = {https://aclanthology.org/2022.lrec-1.785}
}

@InProceedings{ito-EtAl:2022:LREC,
  author   = {Ito, Koichiro and Murata, Masaki and Ohno, Tomohiro and Matsubara, Shigeki},
  title   = {Construction of Responsive Utterance Corpus for Attentive Listening Response Production},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7244--7252},
  abstract = {In Japan, the number of single-person households, particularly among the elderly, is increasing. Consequently, opportunities for people to narrate are being reduced. To address this issue, conversational agents, e.g., communication robots and smart speakers, are expected to play the role of the listener. To realize these agents, this paper describes the collection of conversational responses by listeners that demonstrate attentive listening attitudes toward narrative speakers, and a method to annotate existing narrative speech with responsive utterances is proposed. To summarize, 148,962 responsive utterances by 11 listeners were collected in a narrative corpus comprising 13,234 utterance units. The collected responsive utterances were analyzed in terms of response frequency, diversity, coverage, and naturalness. These results demonstrated that diverse and natural responsive utterances were collected by the proposed method in an efficient and comprehensive manner. To demonstrate the practical use of the collected responsive utterances, an experiment was conducted, in which response generation timings were detected in narratives.},
  url   = {https://aclanthology.org/2022.lrec-1.786}
}

@InProceedings{song-EtAl:2022:LREC,
  author   = {Song, Christopher and Harwath, David and Alhanai, Tuka and Glass, James},
  title   = {Speak: A Toolkit Using Amazon Mechanical Turk to Collect and Validate Speech Audio Recordings},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7253--7258},
  abstract = {We present Speak, a toolkit that allows researchers to crowdsource speech audio recordings using Amazon Mechanical Turk (MTurk). Speak allows MTurk workers to submit speech recordings in response to a task prompt and stimulus (e.g. image, text excerpt, audio file) defined by researchers, a functionality that is not

natively offered by MTurk at the time of writing this paper. Importantly, the toolkit employs numerous measures to ensure that speech recordings collected are of adequate quality, in order to avoid accepting unusable data and prevent abuse/fraud. Speak has demonstrated utility, having collected over 600,000 recordings to date. The toolkit is open-source and available for download.},
  url   = {https://aclanthology.org/2022.lrec-1.787}
}

@InProceedings{lovenia-EtAl:2022:LREC,
  author   = {Lovenia, Holy and Cahyawijaya, Samuel and Winata, Genta and Xu, Peng and Xu, Yan and Liu, Zihan and Frieske, Rita and Yu, Tiezheng and Dai, Wenliang and Barezi, Elham J. and Chen, Qifeng and Ma, Xiaojuan and Shi, Bertram and Fung, Pascale},
  title   = {ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7259--7268},
  abstract = {Code-switching is a speech phenomenon occurring when a speaker switches language during a conversation. Despite the spontaneous nature of code-switching in conversational spoken language, most existing works collect code-switching data from read speech instead of spontaneous speech. ASCEND (A Spontaneous Chinese-English Dataset) is a high-quality Mandarin Chinese-English code-switching corpus built on spontaneous multi-turn conversational dialogue sources collected in Hong Kong. We report ASCEND's design and procedure for collecting the speech data, including annotations. ASCEND consists of 10.62 hours of clean speech, collected from 23 bilingual speakers of Chinese and English. Furthermore, we conduct baseline experiments using pre-trained wav2vec 2.0 models, achieving a best performance of 22.69\% character error rate and 27.05\% mixed error rate.},
  url   = {https://aclanthology.org/2022.lrec-1.788}
}

@InProceedings{altamimi-EtAl:2022:LREC,
  author   = {Al-Tamimi, Jalal and Schiel, Florian and Khattab, Ghada and Sokhey, Navdeep and Amazouz, Djegdjiga and Dallak, Abdulrahman and Moussa, Hajar},
  title   = {A Romanization System and WebMAUS Aligner for Arabic Varieties},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7269--7276},

  abstract = {This paper presents the results of an ongoing collaboration to develop an Arabic variety-independent romanization system that aims to homogenize and simplify the romanization of the Arabic script, and introduces an Arabic variety-independent WebMAUS service offering a free to use forced-alignment service fully integrated within the WebMAUS services. We present the rationale for developing such a system, highlighting the need for a detailed romanization system with graphemes corresponding to the phonemic short and long vowels/consonants in Arabic varieties. We describe how the acoustic model was created, followed by several hands-on recipes for applying the forced alignment webservice either online or programatically. Finally, we discuss some of the issues we faced during the development of the system.},
  url   = {https://aclanthology.org/2022.lrec-1.789}
}

@InProceedings{eder-EtAl:2022:LREC,
  author   = {Eder, Elisabeth and Wiegand, Michael and Krieg-Holz, Ulrike and Hahn, Udo},
  title   = {"Beste Grüße, Maria Meyer" — Pseudonymization of Privacy-Sensitive Information in Emails},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {741--752},
  abstract = {The exploding amount of user-generated content has spurred NLP research to deal with documents from various digital communication formats (tweets, chats, emails, etc.). Using these texts as language resources implies complying with legal data privacy regulations. To protect the personal data of individuals and preclude their identification, we employ pseudonymization. More precisely, we identify those text spans that carry information revealing an individual's identity (e.g., names of persons, locations, phone numbers, or dates) and subsequently substitute them with synthetically generated surrogates. Based on CodE Alltag, a German-language email corpus, we address two tasks. The first task is to evaluate various architectures for the automatic recognition of privacy-sensitive entities in raw data. The second task examines the applicability of pseudonymized data as training data for such systems since models learned on original data cannot be published for reasons of privacy protection. As outputs of both tasks, we, first, generate a new pseudonymized version of CodE Alltag compliant with the legal requirements of the General Data Protection Regulation (GDPR). Second, we make accessible a tagger for recognizing privacy-sensitive information in German emails and similar text genres, which is trained on already pseudonymized data.},
  url   = {https://aclanthology.org/2022.lrec-1.79}
}

@InProceedings{sikasote-anastasopoulos:2022:LREC,

  author   = {Sikasote, Claytone and Anastasopoulos, Antonios},
  title   = {BembaSpeech: A Speech Recognition Corpus for the Bemba Language},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7277--7283},
  abstract = {We present a preprocessed, ready-to-use automatic speech recognition corpus, BembaSpeech, consisting over 24 hours of read speech in the Bemba language, a written but low-resourced language spoken by over 30\% of the population in Zambia. To assess its usefulness for training and testing ASR systems for Bemba, we explored different approaches; supervised pre-training (training from scratch), cross-lingual transfer learning from a monolingual English pre-trained model using DeepSpeech on the portion of the dataset and fine-tuning large scale self-supervised Wav2Vec2.0 based multilingual pre-trained models on the complete BembaSpeech corpus. From our experiments, the 1 billion XLS-R parameter model gives the best results. The model achieves a word error rate (WER) of 32.91\%, results demonstrating that model capacity significantly improves performance and that multilingual pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained English model on the BembaSpeech for the Bemba ASR. Lastly, results also show that the corpus can be used for building ASR systems for Bemba language.},
  url   = {https://aclanthology.org/2022.lrec-1.790}
}

@InProceedings{lai-EtAl:2022:LREC,
  author   = {Lai, Viet and Pouran Ben Veyseh, Amir and Dernoncourt, Franck and Nguyen, Thien Huu},
  title   = {BehanceCC: A ChitChat Detection Dataset For Livestreaming Video Transcripts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7284--7290},
  abstract = {Livestreaming videos have become an effective broadcasting method for both video sharing and educational purposes. However, livestreaming videos contain a considerable amount of off-topic content (i.e., up to 50\%) which introduces significant noises and data load to downstream applications. This paper presents BehanceCC, a new human-annotated benchmark dataset for off-topic detection (also called chitchat detection) in livestreaming video transcripts. In addition to describing the challenges of the dataset, our extensive experiments of various baselines reveal the complexity of chitchat detection for livestreaming videos and suggest potential future research directions for this task. The

dataset will be made publicly available to foster research in this area.},
  url   = {https://aclanthology.org/2022.lrec-1.791}
}

@InProceedings{li-EtAl:2022:LREC4,
  author   = {Li, Sheng and Li, Jiyi and Liu, Qianying and Gong, Zhuo},
  title   = {Adversarial Speech Generation and Natural Speech Recovery for Speech Content Protection},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7291--7297},
  abstract = {With the advent of the General Data Protection Regulation (GDPR) and increasing privacy concerns, the sharing of speech data is faced with significant challenges. Protecting the sensitive content of speech is the same important as the voiceprint. This paper proposes an effective speech content protection method by constructing a frame-by-frame adversarial speech generation system. We revisited the adversarial examples generating method in the recent machine learning field and selected the phonetic state sequence of sensitive speech for the adversarial examples generation. We build an adversarial speech collection. Moreover, based on the speech collection, we proposed a neural network-based frame-by-frame mapping method to recover the speech content by converting from the adversarial speech to the human speech. Experiment shows our proposed method can encode and recover any sensitive audio, and our method is easy to be conducted with publicly available resources of speech recognition technology.},
  url   = {https://aclanthology.org/2022.lrec-1.792}
}

@InProceedings{forj-EtAl:2022:LREC,
  author   = {Forjó, Maria and Neto, Daniel and Abad, Alberto and Pinto, HSofia and Gago, Joaquim},
  title   = {A new European Portuguese corpus for the study of Psychosis through speech analysis},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7298--7304},
  abstract = {Psychosis is a clinical syndrome characterized by the presence of symptoms such as hallucinations, thought disorder and disorganized speech. Several studies have used machine learning, combined with speech and natural language processing methods to aid in the diagnosis process of this disease. This paper describes the creation of the first European Portuguese corpus for the

identification of the presence of speech characteristics of psychosis, which contains samples of 92 participants, 56 controls and 36 individuals diagnosed with psychosis and medicated. The corpus was used in a set of experiments that allowed identifying the most promising feature set to perform the classification: the combination of acoustic and speech metric features. Several classifiers were implemented to study which ones entailed the best performance depending on the task and feature set. The most promising results obtained for the entire corpus were achieved when identifying individuals with a Multi-Layer Perceptron classifier and reached an 87.5\% accuracy. Focusing on the gender dependent results, the overall best results were 90.9\% and 82.9\% accuracy, for female and male subjects respectively. Lastly, the experiments performed lead us to conjecture that spontaneous speech presents more identifiable characteristics than read speech to differentiate healthy and patients diagnosed with psychosis.},
  url   = {https://aclanthology.org/2022.lrec-1.793}
}

@InProceedings{sini-EtAl:2022:LREC,
  author   = {SINI, Aghilas and Lolive, Damien and Barbot, Nelly and Alain, Pierre},
  title   = {Investigating Inter- and Intra-speaker Voice Conversion using Audiobooks},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7305--7313},
  abstract = {Audiobook readers play with their voices to emphasize some text passages, highlight discourse changes or significant events, or in order to make listening easier and entertaining. A dialog is a central passage in audiobooks where the reader applies significant voice transformation, mainly prosodic modifications, to realize character properties and changes. However, these intra-speaker modifications are hard to reproduce with simple text-to-speech synthesis. The manner of vocalizing characters involved in a given story depends on the text style and differs from one speaker to another. In this work, this problem is investigated through the prism of voice conversion. We propose to explore modifying the narrator’s voice to fit the context of the story, such as the character who is speaking, using voice conversion. To this end, two complementary experiments are designed: the first one aims to assess the quality of our Phonetic PosteriorGrams (PPG)-based voice conversion system using parallel data. Subjective evaluations with naive raters are conducted to estimate the quality of the signal generated and the speaker similarity. The second experiment applies an intra-speaker voice conversion, considering narration passages and direct speech passages as two distinct speakers. Data are then nonparallel and the dissimilarity between character and narrator is subjectively measured.},
  url   = {https://aclanthology.org/2022.lrec-1.794}

}

@InProceedings{rolland-EtAl:2022:LREC,
  author   = {Rolland, Thomas and Abad, Alberto and Cucchiarini, Catia and Strik, Helmer},
  title   = {Multilingual Transfer Learning for Children Automatic Speech Recognition},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7314--7320},
  abstract = {Despite recent advances in automatic speech recognition (ASR), the recognition of children's speech still remains a significant challenge. This is mainly due to the high acoustic variability and the limited amount of available training data. The latter problem is particularly evident in languages other than English, which are usually less-resourced. In the current paper, we address children ASR in a number of less-resourced languages by combining several small-sized children speech corpora from these languages. In particular, we address the following research question: Does a novel two-step training strategy in which multilingual learning is followed by language-specific transfer learning outperform conventional single language/task training for children speech, as well as multilingual and transfer learning alone? Based on previous experimental results with English, we hypothesize that multilingual learning provides a better generalization of the underlying characteristics of children’s speech. Our results provide a positive answer to our research question, by showing that using transfer learning on top of a multilingual model for an unseen language outperforms conventional single language-specific learning.},
  url   = {https://aclanthology.org/2022.lrec-1.795}
}

@InProceedings{pouranbenveyseh-EtAl:2022:LREC,
  author   = {Pouran Ben Veyseh, Amir and Lai, Viet and Dernoncourt, Franck and Nguyen, Thien Huu},
  title   = {BehanceQA: A New Dataset for Identifying Question-Answer Pairs in Video Transcripts},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7321--7327},
  abstract = {Question-Answer (QA) is one of the effective methods for storing knowledge which can be used for future retrieval. As such, identifying mentions of questions and their answers in text is necessary for a knowledge construction and retrieval systems. In the literature, QA identification has been well studied in the NLP

community. However, most of the prior works are restricted to formal written documents such as papers or websites. As such, Questions and Answers that are presented in informal/noisy documents have not been adequately studied. One of the domains that can significantly benefit from QA identification is the domain of livestreaming video transcripts that involve abundant QA pairs to provide valuable knowledge for future users and services. Since video transcripts are often transcribed automatically for scale, they are prone to errors. Combined with the informal nature of discussion in a video, prior QA identification systems might not be able to perform well in this domain. To enable comprehensive research in this domain, we present a large-scale QA identification dataset annotated by human over transcripts of 500 hours of streamed videos. We employ Behance.net to collect the videos and their automatically obtained transcripts. Furthermore, we conduct extensive analysis on the annotated dataset to understand the complexity of QA identification for livestreaming video transcripts. Our experiments show that the annotated dataset presents unique challenges for existing methods and more research is necessary to explore more effective methods. The dataset and the models developed in this work will be publicly released for future research.},
  url   = {https://aclanthology.org/2022.lrec-1.796}
}

@InProceedings{dafnis-EtAl:2022:LREC,
  author   = {Dafnis, Konstantinos M. and Chroni, Evgenia and Neidle, Carol and Metaxas, Dimitri},
  title   = {Bidirectional Skeleton-Based Isolated Sign Recognition using Graph Convolutional Networks},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7328--7338},
  abstract = {To improve computer-based recognition from video of isolated signs from American Sign Language (ASL), we propose a new skeleton-based method that involves explicit detection of the start and end frames of signs, trained on the ASLLVD dataset; it uses linguistically relevant parameters based on the skeleton input. Our method employs a bidirectional learning approach within a Graph Convolutional Network (GCN) framework. We apply this method to the WLASL dataset, but with corrections to the gloss labeling to ensure consistency in the labels assigned to different signs; it is important to have a 1-1 correspondence between signs and text-based gloss labels. We achieve a success rate of 77.43\% for top-1 and 94.54\% for top-5 using this modified WLASL dataset. Our method, which does not require multi-modal data input, outperforms other state-of-the-art approaches on the same modified WLASL dataset, demonstrating the importance of both attention to the start and end frames of signs and the use of bidirectional data streams in the GCNs for isolated sign recognition.},
  url   = {https://aclanthology.org/2022.lrec-1.797}

}

@InProceedings{kang-alam-fathan:2022:LREC,
  author   = {Kang, Woohyun and Alam, Md Jahangir and Fathan, Abderrahim},
  title   = {Deep learning-based end-to-end spoken language identification system for domain-mismatched scenario},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7339--7343},
  abstract = {Domain mismatch is a critical issue when it comes to spoken language identification. To overcome the domain mismatch problem, we have applied several architectures and deep learning strategies which have shown good results in cross-domain speaker verification tasks to spoken language identification. Our systems were evaluated on the Oriental Language Recognition (OLR) Challenge 2021 Task 1 dataset, which provides a set of cross-domain language identification trials. Among our experimented systems, the best performance was achieved by using the mel frequency cepstral coefficient (MFCC) and pitch features as input and training the ECAPA-TDNN system with a flow-based regularization technique, which resulted in a Cavg of 0.0631 on the OLR 2021 progress set.},
  url   = {https://aclanthology.org/2022.lrec-1.798}
}

@InProceedings{kitagawa-leow-nishizaki:2022:LREC,
  author   = {Kitagawa, Tomoki and Leow, Chee Siang and Nishizaki, Hiromitsu},
  title   = {Handwritten Character Generation using Y-Autoencoder for Character Recognition Model Training},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7344--7351},
  abstract = {It is well-known that the deep learning-based optical character recognition (OCR) system needs a large amount of data to train a high-performance character recognizer. However, it is costly to collect a large amount of realistic handwritten characters. This paper introduces a Y-Autoencoder (Y-AE)-based handwritten character generator to generate multiple Japanese Hiragana characters with a single image to increase the amount of data for training a handwritten character recognizer. The adaptive instance normalization (AdaIN) layer allows the generator to be trained and generate handwritten character images without paired-character image labels. The experiment shows that the Y-AE could generate Japanese character images then used to train the handwritten character recognizer, producing an F1-score improved from 0.8664 to 0.9281. We

further analyzed the usefulness of the Y-AE-based generator with shape images, out-of-character (OOC) images, which have different character images styles in model training. The result showed that the generator could generate a handwritten image with a similar style to that of the input character.},
  url   = {https://aclanthology.org/2022.lrec-1.799}
}

@InProceedings{sio-morgadodacosta:2022:LREC,
  author   = {Sio, Ut Seong and Morgado da Costa, Luís},
  title   = {Enriching Linguistic Representation in the Cantonese Wordnet and Building the New Cantonese Wordnet Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {70--78},
  abstract = {This paper reports on the most recent improvements on the Cantonese Wordnet, a wordnet project started in 2019 (Sio and Morgado da Costa, 2019) with the aim of capturing and organizing lexico-semantic information of Hong Kong Cantonese. The improvements we present here extend both the breadth and depth of the Cantonese Wordnet: increasing the general coverage, adding functional categories, enriching verbal representations, as well as creating the Cantonese Wordnet Corpus – a corpus of handcrafted examples where individual senses are shown in context.},
  url   = {https://aclanthology.org/2022.lrec-1.8}
}

@InProceedings{schmeissernieto-nofre-taul:2022:LREC,
  author   = {Schmeisser-Nieto, Wolfgang and Nofre, Montserrat and Taulé, Mariona},
  title   = {Criteria for the Annotation of Implicit Stereotypes},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {753--762},
  abstract = {The growth of social media has brought with it a massive channel for spreading and reinforcing stereotypes. This issue becomes critical when the affected targets are minority groups such as women, the LGBT+ community and immigrants. Although from the perspective of computational linguistics, the detection of this kind of stereotypes is steadily improving, most stereotypes are expressed implicitly and identifying them automatically remains a challenge. One of the problems we found for tackling this issue is the lack of an operationalised definition of implicit stereotypes that would allow us to annotate consistently new corpora by characterising the different forms in which stereotypes appear. In this paper, we present thirteen criteria for annotating implicitness which were

elaborated to facilitate the subjective task of identifying the presence of stereotypes. We also present NewsCom-Implicitness, a corpus of 1,911 sentences, of which 426 comprise explicit and implicit racial stereotypes. An experiment was carried out to evaluate the applicability of these criteria. The results indicate that different criteria obtain different inter-annotator agreement values and that there is a greater agreement when more criteria can be identified in one sentence.},
  url   = {https://aclanthology.org/2022.lrec-1.80}
}

@InProceedings{kanashiropereira:2022:LREC,
  author   = {Kanashiro Pereira, Lis},
  title   = {Attention is All you Need for Robust Temporal Reasoning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7352--7359},
  abstract = {We propose an enhanced adversarial training algorithm for fine-tuning transformer-based language models (i.e., RoBERTa) and apply it to the temporal reasoning task. Current adversarial training approaches for NLP add the adversarial perturbation only to the embedding layer, ignoring the other layers of the model, which might limit the generalization power of adversarial training. Instead, our algorithm searches for the best combination of layers to add the adversarial perturbation. We add the adversarial perturbation to multiple hidden states or attention representations of the model layers. Adding the perturbation to the attention representations performed best in our experiments. Our model can improve performance on several temporal reasoning benchmarks, and establishes new state-of-the-art results.},
  url   = {https://aclanthology.org/2022.lrec-1.800}
}

@InProceedings{kawintiranon-singh:2022:LREC,
  author   = {Kawintiranon, Kornraphop and Singh, Lisa},
  title   = {PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7360--7367},
  abstract = {Transformer-based models have become the state-of-the-art for numerous natural language processing (NLP) tasks, especially for noisy data sets, including social media posts. For example, BERTweet, pre-trained RoBERTa on a large amount of Twitter data, has achieved state-of-the-art results on several Twitter NLP

tasks. We argue that it is not only important to have general pre-trained models for a social media platform, but also domain-specific ones that better capture domain-specific language context. Domain-specific resources are not only important for NLP tasks associated with a specific domain, but they are also useful for understanding language differences across domains. One domain that receives a large amount of attention is politics, more specifically political elections. Towards that end, we release PoliBERTweet, a pre-trained language model trained from BERTweet on over 83M US 2020 election-related English tweets. While the construction of the resource is fairly straightforward, we believe that it can be used for many important downstream tasks involving language, including political misinformation analysis and election public opinion analysis. To show the value of this resource, we evaluate PoliBERTweet on different NLP tasks. The results show that our model outperforms general-purpose language models in domain-specific contexts, highlighting the value of domain-specific models for more detailed linguistic analysis. We also extend other existing language models with a sample of these data and show their value for presidential candidate stance detection, a context-specific task. We release PoliBERTweet and these other models to the community to advance interdisciplinary research related to Election 2020.},
  url   = {https://aclanthology.org/2022.lrec-1.801}
}

@InProceedings{stenger-EtAl:2022:LREC,
  author   = {Stenger, Irina and Georgis, Philip and
Avgustinova, Tania and Möbius, Bernd and Klakow, Dietrich},
  title   = {Modeling the Impact of Syntactic Distance and
Surprisal on Cross-Slavic Text Comprehension},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7368--7376},
  abstract = {We focus on the syntactic variation and measure
syntactic distances between nine Slavic languages (Belarusian,
Bulgarian, Croatian, Czech, Polish, Slovak, Slovene, Russian, and
Ukrainian) using symmetric measures of insertion, deletion and
movement of syntactic units in the parallel sentences of the fable
“The North Wind and the Sun”. Additionally, we investigate phonetic
and orthographic asymmetries between selected languages by means of
the information theoretical notion of surprisal. Syntactic distance
and surprisal are, thus, considered as potential predictors of
mutual intelligibility between related languages. In spoken and
written cloze test experiments for Slavic native speakers, the
presented predictors will be validated as to whether variations in
syntax lead to a slower or impeded intercomprehension of Slavic
texts.},
  url   = {https://aclanthology.org/2022.lrec-1.802}
}

@InProceedings{dhananjaya-EtAl:2022:LREC,
  author   = {Dhananjaya, Vinura and Demotte, Piyumal and
Ranathunga, Surangika and Jayasena, Sanath},
  title   = {BERTifying Sinhala - A Comprehensive Analysis of Pre-
trained Language Models for Sinhala Text Classification},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7377--7385},
  abstract = {This research provides the first comprehensive
analysis of the performance of pre-trained language models for
Sinhala text classification. We test on a set of different Sinhala
text classification tasks and our analysis shows that out of the
pre-trained multilingual models that include Sinhala (XLM-R, LaBSE,
and LASER), XLM-R is the best model by far for Sinhala text
classification. We also pre-train two RoBERTa-based monolingual
Sinhala models, which are far superior to the existing pre-trained
language models for Sinhala. We show that when fine-tuned, these
pre-trained language models set a very strong baseline for Sinhala
text classification and are robust in situations where labeled data
is insufficient for fine-tuning. We further provide a set of
recommendations for using pre-trained models for Sinhala text
classification. We also introduce new annotated datasets useful for
future research in Sinhala text classification and publicly release
our pre-trained models.},
  url   = {https://aclanthology.org/2022.lrec-1.803}
}

@InProceedings{daason-loftsson:2022:LREC,
  author   = {Daðason, Jón Friðrik and Loftsson, Hrafn},
  title   = {Pre-training and Evaluating Transformer-based
Language Models for Icelandic},
  booktitle   = {Proceedings of the Language Resources and
Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {7386--7391},
  abstract = {In this paper, we evaluate several Transformer-based
language models for Icelandic on four downstream tasks: Part-of-
Speech tagging, Named Entity Recognition. Dependency Parsing, and
Automatic Text Summarization. We pre-train four types of monolingual
ELECTRA and ConvBERT models and compare our results to a previously
trained monolingual RoBERTa model and the multilingual mBERT model.
We find that the Transformer models obtain better results, often by
a large margin, compared to previous state-of-the-art models.
Furthermore, our results indicate that pre-training larger language
models results in a significant reduction in error rates in
comparison to smaller models. Finally, our results show that the
monolingual models for Icelandic outperform a comparably sized

multilingual model.},
  url   = {https://aclanthology.org/2022.lrec-1.804}
}

@InProceedings{klumpp-EtAl:2022:LREC,
  author   = {Klumpp, Philipp and Arias, Tomas and Pérez-Toro, Paula Andrea and Noeth, Elmar and Orozco-Arroyave, Juan},
  title   = {Common Phone: A Multilingual Dataset for Robust Acoustic Modelling},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {763--768},
  abstract = {Current state of the art acoustic models can easily comprise more than 100 million parameters. This growing complexity demands larger training datasets to maintain a decent generalization of the final decision function. An ideal dataset is not necessarily large in size, but large with respect to the amount of unique speakers, utilized hardware and varying recording conditions. This enables a machine learning model to explore as much of the domain-specific input space as possible during parameter estimation. This work introduces Common Phone, a gender-balanced, multilingual corpus recorded from more than 76.000 contributors via Mozilla's Common Voice project. It comprises around 116 hours of speech enriched with automatically generated phonetic segmentation. A Wav2Vec 2.0 acoustic model was trained with the Common Phone to perform phonetic symbol recognition and validate the quality of the generated phonetic annotation. The architecture achieved a PER of 18.1 \% on the entire test set, computed with all 101 unique phonetic symbols, showing slight differences between the individual languages. We conclude that Common Phone provides sufficient variability and reliable phonetic annotation to help bridging the gap between research and application of acoustic models.},
  url   = {https://aclanthology.org/2022.lrec-1.81}
}

@InProceedings{alhaff-EtAl:2022:LREC,
  author   = {Al-Haff, Karim and Jarrar, Mustafa and Hammouda, Tymaa and Zaraket, Fadi},
  title   = {Curras + Baladi: Towards a Levantine Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {769--778},
  abstract = {This paper presents two-fold contributions: a full revision of the Palestinian morphologically annotated corpus (Curras), and a newly annotated Lebanese corpus (Baladi). Both corpora can be used as a more general Levantine corpus. Baladi

consists of around 9.6K morphologically annotated tokens. Each token was manually annotated with several morphological features and using LDC’s SAMA lemmas and tags. The inter-annotator evaluation on most features illustrates 78.5\% Kappa and 90.1\% F1-Score. Curras was revised by refining all annotations for accuracy, normalization and unification of POS tags, and linking with SAMA lemmas. This revision was also important to ensure that both corpora are compatible and can help to bridge the nuanced linguistic gaps that exist between the two highly mutually intelligible dialects. Both corpora are publicly available through a web portal.},
  url   = {https://aclanthology.org/2022.lrec-1.82}
}

@InProceedings{yamada-EtAl:2022:LREC,
  author   = {Yamada, Hiroaki and Tokunaga, Takenobu and Ohara, Ryutaro and Takeshita, Keisuke and Sumida, Mihoko},
  title   = {Annotation Study of Japanese Judgments on Tort for Legal Judgment Prediction with Rationales},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {779--790},
  abstract = {This paper describes a comprehensive annotation study on Japanese judgment documents in civil cases. We aim to build an annotated corpus designed for Legal Judgment Prediction (LJP), especially for torts. Our annotation scheme contains annotations of whether tort is accepted by judges as well as its corresponding rationales for explainability purpose. Our annotation scheme extracts decisions and rationales at character-level. Moreover, the scheme can capture the explicit causal relation between judge's decisions and their corresponding rationales, allowing multiple decisions in a document. To obtain high-quality annotation, we developed an annotation scheme with legal experts, and confirmed its reliability by agreement studies with Krippendorff's alpha metric. The result of the annotation study suggests the proposed annotation scheme can produce a dataset of Japanese LJP at reasonable reliability.},
  url   = {https://aclanthology.org/2022.lrec-1.83}
}

@InProceedings{ruiter-EtAl:2022:LREC,
  author   = {Ruiter, Dana and Reiners, Liane and D'Sa, Ashwin Geet and Kleinbauer, Thomas and Fohr, Dominique and Illina, Irina and Klakow, Dietrich and Schemer, Christian and Monnier, Angeliki},
  title   = {Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},

  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {791--804},
  abstract = {Even though hate speech (HS) online has been an important object of research in the last decade, most HS-related corpora over-simplify the phenomenon of hate by attempting to label user comments as "hate" or "neutral". This ignores the complex and subjective nature of HS, which limits the real-life applicability of classifiers trained on these corpora. In this study, we present the M-Phasis corpus, a corpus of ~9k German and French user comments collected from migration-related news articles. It goes beyond the "hate"-"neutral" dichotomy and is instead annotated with 23 features, which in combination become descriptors of various types of speech, ranging from critical comments to implicit and explicit expressions of hate. The annotations are performed by 4 native speakers per language and achieve high (0.77 <= k <= 1) inter-annotator agreements. Besides describing the corpus creation and presenting insights from a content, error and domain analysis, we explore its data characteristics by training several classification baselines.},
  url   = {https://aclanthology.org/2022.lrec-1.84}
}

@InProceedings{lapshinovakoltunski-EtAl:2022:LREC,
  author   = {Lapshinova-Koltunski, Ekaterina and Ferreira, Pedro Augusto and Lartaud, Elina and Hardmeier, Christian},
  title   = {ParCorFull2.0: a Parallel Corpus Annotated with Full Coreference},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {805--813},
  abstract = {In this paper, we describe ParCorFull2.0, a parallel corpus annotated with full coreference chains for multiple languages, which is an extension of the existing corpus ParCorFull (Lapshinova-Koltunski et al., 2018). Similar to the previous version, this corpus has been created to address translation of coreference across languages, a phenomenon still challenging for machine translation (MT) and other multilingual natural language processing (NLP) applications. The current version of the corpus that we present here contains not only parallel texts for the language pair English-German, but also for English-French and English-Portuguese, which are all major European languages. The new language pairs belong to the Romance languages. The addition of a new language group creates a need of extension not only in terms of texts added, but also in terms of the annotation guidelines. Both French and Portuguese contain structures not found in English and German. Moreover, Portuguese is a pro-drop language bringing even more systemic differences in the realisation of coreference into our cross-lingual resources. These differences cause problems for multilingual coreference resolution and machine translation. Our

parallel corpus with full annotation of coreference will be a valuable resource with a variety of uses not only for NLP applications, but also for contrastive linguists and researchers in translation studies.},
  url   = {https://aclanthology.org/2022.lrec-1.85}
}

@InProceedings{boritchev-amblard:2022:LREC,
  author   = {Boritchev, Maria and Amblard, Maxime},
  title   = {A Multi-Party Dialogue Ressource in French},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {814--823},
  abstract = {We presentDialogues in Games(DinG), a corpus of manual transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of the board game Catan. Our objective is to make available a quality resource for French, composed of long dialogues, to facilitate their study in the style of (Asher et al., 2016). In a general dialogue setting, participants share personal information, which makes it impossible to disseminate the resource freely and openly. In DinG, the attention of the participants is focused on the game, which prevents them from talking about themselves. In addition, we are conducting a study on the nature of the questions in dialogue, through annotation (Cruz Blandon et al., 2019), in order to develop more natural automatic dialogue systems},
  url   = {https://aclanthology.org/2022.lrec-1.86}
}

@InProceedings{zaragozabernabeu-EtAl:2022:LREC,
  author   = {Zaragoza-Bernabeu, Jaume and Ramírez-Sánchez, Gema and Bañón, Marta and Ortiz Rojas, Sergio},
  title   = {Bicleaner AI: Bicleaner Goes Neural},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {824--831},
  abstract = {This paper describes the experiments carried out during the development of the latest version of Bicleaner, named Bicleaner AI, a tool that aims at detecting noisy sentences in parallel corpora. The tool, which now implements a new neural classifier, uses state-of-the-art techniques based on pre-trained transformer-based language models fine-tuned on a binary classification task. After that, parallel corpus filtering is performed, discarding the sentences that have lower probability of being mutual translations. Our experiments, based on the training of neural machine translation (NMT) with corpora filtered using

Bicleaner AI for two different scenarios, show significant improvements in translation quality compared to the previous version of the tool which implemented a classifier based on Extremely Randomized Trees.},
  url   = {https://aclanthology.org/2022.lrec-1.87}
}

@InProceedings{katinskaia-EtAl:2022:LREC,
  author   = {Katinskaia, Anisia and Lebedeva, Maria and Hou, Jue and Yangarber, Roman},
  title   = {Semi-automatically Annotated Learner Corpus for Russian},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {832--839},
  abstract = {We present ReLCo--- the Revita Learner Corpus---a new semi-automatically annotated learner corpus for Russian. The corpus was collected while several thousand L2 learners were performing exercises using the Revita language-learning system. All errors were detected automatically by the system and annotated by type. Part of the corpus was annotated manually---this part was created for further experiments on automatic assessment of grammatical correctness. The Learner Corpus provides valuable data for studying patterns of grammatical errors, experimenting with grammatical error detection and grammatical error correction, and developing new exercises for language learners. Automating the collection and annotation makes the process of building the learner corpus much cheaper and faster, in contrast to the traditional approach of building learner corpora. We make the data publicly available.},
  url   = {https://aclanthology.org/2022.lrec-1.88}
}

@InProceedings{batsuren-EtAl:2022:LREC,
  author   = {Batsuren, Khuyagbaatar and Goldman, Omer and Khalifa, Salam and Habash, Nizar and Kieraś, Witold and Bella, Gábor and Leonard, Brian and Nicolai, Garrett and Gorman, Kyle and Ate, Yustinus Ghanggo and Ryskina, Maria and Mielke, Sabrina and Budianskaya, Elena and El-Khaissi, Charbel and Pimentel, Tiago and Gasser, Michael and Lane, William Abbott and Raj, Mohit and Coler, Matt and Samame, Jaime Rafael Montoya and Camaiteri, Delio Siticonatzi and Rojas, Esaú Zumaeta and Francis, Didier López and Oncevay, Arturo and Bautista, Juan López and Villegas, Gema Celeste Silva and Hennigen, Lucas Torroba and Ek, Adam and Guriel, David and Dirix, Peter and Bernardy, Jean-Philippe and Scherbakov, Andrey and Bayyr-ool, Aziyana and Anastasopoulos, Antonios and Zariquiey, Roberto and Sheifer, Karina and Ganieva, Sofya and Cruz, Hilaria and Karahóǧa, Ritván and Markantonatou, Stella and Pavlidis, George and Plugaryov, Matvey and Klyachko, Elena and Salehi, Ali and Angulo, Candy and Baxi, Jatayu and Krizhanovsky, Andrew and

Krizhanovskaya, Natalia and Salesky, Elizabeth and Vania, Clara and Ivanova, Sardana and White, Jennifer and Maudslay, Rowan Hall and Valvoda, Josef and Zmigrod, Ran and Czarnowska, Paula and Nikkarinen, Irene and Salchak, Aelita and bhatt, brijesh and Straughn, Christopher and Liu, Zoey and Washington, Jonathan North and Pinter, Yuval and Ataman, Duygu and Wolinski, Marcin and Suhardijanto, Totok and Yablonskaya, Anna and Stoehr, Niklas and Dolatian, Hossep and Nuriah, Zahroh and Ratan, Shyam and Tyers, Francis M. and Ponti, Edoardo M. and Aiton, Grant and Arora, Aryaman and Hatcher, Richard J. and Kumar, Ritesh and Young, Jeremiah and Rodionova, Daria and Yemelina, Anastasia and Andrushko, Taras and Marchenko, Igor and Mashkovtseva, Polina and Serova, Alexandra and Prud'hommeaux, Emily and Nepomniashchaya, Maria and giunchiglia, fausto and Chodroff, Eleanor and Hulden, Mans and Silfverberg, Miikka and McCarthy, Arya D. and Yarowsky, David and Cotterell, Ryan and Tsarfaty, Reut and Vylomova, Ekaterina},
  title = {UniMorph 4.0: Universal Morphology},
  booktitle = {Proceedings of the Language Resources and Evaluation Conference},
  month = {June},
  year = {2022},
  address = {Marseille, France},
  publisher = {European Language Resources Association},
  pages = {840--855},
  abstract = {The Universal Morphology (UniMorph) project is a collaborative effort providing broad-coverage instantiated normalized morphological inflection tables 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. This paper presents the expansions and improvements on several fronts that were made in the last couple of years (since McCarthy et al. (2020)). Collaborative efforts by numerous linguists have added 66 new languages, including 24 endangered languages. We have implemented several improvements to the extraction pipeline to tackle some issues, e.g., missing gender and macrons information. We have amended the schema to use a hierarchical structure that is needed for morphological phenomena like multiple-argument agreement and case stacking, while adding some missing morphological features to make the schema more inclusive.In light of the last UniMorph release, we also augmented the database with morpheme segmentation for 16 languages. Lastly, this new release makes a push towards inclusion of derivational morphology in UniMorph by enriching the data and annotation schema with instances representing derivational processes from MorphyNet.},
  url = {https://aclanthology.org/2022.lrec-1.89}
}

@InProceedings{habash-palfreyman:2022:LREC,
  author = {Habash, Nizar and Palfreyman, David},
  title = {ZAEBUC: An Annotated Arabic-English Bilingual Writer Corpus},
  booktitle = {Proceedings of the Language Resources and

Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {79--88},
  abstract = {We present ZAEBUC, an annotated Arabic-English bilingual writer corpus comprising short essays by first-year university students at Zayed University in the United Arab Emirates. We describe and discuss the various guidelines and pipeline processes we followed to create the annotations and quality check them. The annotations include spelling and grammar correction, morphological tokenization, Part-of-Speech tagging, lemmatization, and Common European Framework of Reference (CEFR) ratings. All of the annotations are done on Arabic and English texts using consistent guidelines as much as possible, with tracked alignments among the different annotations, and to the original raw texts. For morphological tokenization, POS tagging, and lemmatization, we use existing automatic annotation tools followed by manual correction. We also present various measurements and correlations with preliminary insights drawn from the data and annotations. The publicly available ZAEBUC corpus and its annotations are intended to be the stepping stones for additional annotations.},
  url   = {https://aclanthology.org/2022.lrec-1.9}
}

@InProceedings{kalpakchi-boye:2022:LREC,
  author   = {Kalpakchi, Dmytro and Boye, Johan},
  title   = {Textinator: an Internationalized Tool for Annotation and Human Evaluation in Natural Language Processing and Generation},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {856--866},
  abstract = {We release an internationalized annotation and human evaluation bundle, called Textinator, along with documentation and video tutorials. Textinator allows annotating data for a wide variety of NLP tasks, and its user interface is offered in multiple languages, lowering the entry threshold for domain experts. The latter is, in fact, quite a rare feature among the annotation tools, that allows controlling for possible unintended biases introduced due to hiring only English-speaking annotators. We illustrate the rarity of this feature by presenting a thorough systematic comparison of Textinator to previously published annotation tools along 9 different axes (with internationalization being one of them). To encourage researchers to design their human evaluation before starting to annotate data, Textinator offers an easy-to-use tool for human evaluations allowing importing surveys with potentially hundreds of evaluation items in one click. We finish by presenting several use cases of annotation and evaluation projects conducted using pre-release versions of Textinator. The presented

use cases do not represent Textinator's full annotation or evaluation capabilities, and interested readers are referred to the online documentation for more information.},
  url   = {https://aclanthology.org/2022.lrec-1.90}
}

@InProceedings{ollagnier-EtAl:2022:LREC,
  author   = {Ollagnier, Anaïs and Cabrio, Elena and Villata, Serena and Blaya, Catherine},
  title   = {CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {867--875},
  abstract = {Over the past decades, the number of episodes of cyber aggression occurring online has grown substantially, especially among teens. Most solutions investigated by the NLP community to curb such online abusive behaviors consist of supervised approaches relying on annotated data extracted from social media. However, recent studies have highlighted that private instant messaging platforms are major mediums of cyber aggression among teens. As such interactions remain invisible due to the app privacy policies, very few datasets collecting aggressive conversations are available for the computational analysis of language. In order to overcome this limitation, in this paper we present the CyberAgressionAdo-V1 dataset, containing aggressive multiparty chats in French collected through a role-playing game in high-schools, and annotated at different layers. We describe the data collection and annotation phases, carried out in the context of a EU and a national research projects, and provide insightful analysis on the different types of aggression and verbal abuse depending on the targeted victims (individuals or communities) emerging from the collected data.},
  url   = {https://aclanthology.org/2022.lrec-1.91}
}

@InProceedings{jahan-oussalah-arhab:2022:LREC,
  author   = {Jahan, Md Saroar and Oussalah, Mourad and Arhab, Nabil},
  title   = {Finnish Hate-Speech Detection on Social Media Using CNN and FinBERT},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {876--882},
  abstract = {There has been a lot of research in identifying hate posts from social media because of their detrimental effects on both

individuals and society. The majority of this research has concentrated on English, although one notices the emergence of multilingual detection tools such as multilingual-BERT (mBERT). However, there is a lack of hate speech datasets compared to English, and a multilingual pre-trained model often contains fewer tokens for other languages. This paper attempts to contribute to hate speech identification in Finnish by constructing a new hate speech dataset that is collected from a popular forum (Suomi24). Furthermore, we have experimented with FinBERT pre-trained model performance for Finnish hate speech detection compared to state-of-the-art mBERT and other practices. In addition, we tested the performance of FinBERT compared to fastText as embedding, which employed with Convolution Neural Network (CNN). Our results showed that FinBERT yields a 91.7\% accuracy and 90.8\% F1 score value, which outperforms all state-of-art models, including multilingual-BERT and CNN.},
  url   = {https://aclanthology.org/2022.lrec-1.92}
}

@InProceedings{moon-EtAl:2022:LREC1,
  author   = {Moon, Hyeonseok and park, chanjun and Lee, Seolhwa and Seo, Jaehyung and Lee, Jungseob and Eo, Sugyeong and Lim, Heuiseok},
  title   = {Empirical Analysis of Noising Scheme based Synthetic Data Generation for Automatic Post-editing},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {883--891},
  abstract = {Automatic post-editing (APE) refers to a research field that aims to automatically correct errors included in the translation sentences derived by the machine translation system. This study has several limitations, considering the data acquisition, because there is no official dataset for most language pairs. Moreover, the amount of data is restricted even for language pairs in which official data has been released, such as WMT. To solve this problem and promote universal APE research regardless of APE data existence, this study proposes a method for automatically generating APE data based on a noising scheme from a parallel corpus. Particularly, we propose a human mimicking errors-based noising scheme that considers a practical correction process at the human level. We propose a precise inspection to attain high performance, and we derived the optimal noising schemes that show substantial effectiveness. Through these, we also demonstrate that depending on the type of noise, the noising scheme-based APE data generation may lead to inferior performance. In addition, we propose a dynamic noise injection strategy that enables the acquisition of a robust error correction capability and demonstrated its effectiveness by comparative analysis. This study enables obtaining a high performance APE model without human-generated data and can promote universal APE research for all language pairs targeting

English.},
  url   = {https://aclanthology.org/2022.lrec-1.93}
}

@InProceedings{edmiston-keung-smith:2022:LREC,
  author   = {Edmiston, Daniel and Keung, Phillip and Smith, Noah A.},
  title   = {Domain Mismatch Doesn't Always Prevent Cross-lingual Transfer Learning},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {892--899},
  abstract = {Cross-lingual transfer learning without labeled target language data or parallel text has been surprisingly effective in zero-shot cross-lingual classification, question answering, unsupervised machine translation, etc. However, some recent publications have claimed that domain mismatch prevents cross-lingual transfer, and their results show that unsupervised bilingual lexicon induction (UBLI) and unsupervised neural machine translation (UNMT) do not work well when the underlying monolingual corpora come from different domains (e.g., French text from Wikipedia but English text from UN proceedings). In this work, we show how a simple initialization regimen can overcome much of the effect of domain mismatch in cross-lingual transfer. We pre-train word and contextual embeddings on the concatenated domain-mismatched corpora, and use these as initializations for three tasks: MUSE UBLI, UN Parallel UNMT, and the SemEval 2017 cross-lingual word similarity task. In all cases, our results challenge the conclusions of prior work by showing that proper initialization can recover a large portion of the losses incurred by domain mismatch.},
  url   = {https://aclanthology.org/2022.lrec-1.94}
}

@InProceedings{papaioannou-EtAl:2022:LREC,
  author   = {Papaioannou, Jens-Michalis and Grundmann, Paul and van Aken, Betty and Samaras, Athanasios and Kyparissidis, Ilias and Giannakoulas, George and Gers, Felix and Loeser, Alexander},
  title   = {Cross-Lingual Knowledge Transfer for Clinical Phenotyping},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {900--909},
  abstract = {Clinical phenotyping enables the automatic extraction of clinical conditions from patient records, which can be beneficial to doctors and clinics worldwide. However, current state-of-the-art

models are mostly applicable to clinical notes written in English. We therefore investigate cross-lingual knowledge transfer strategies to execute this task for clinics that do not use the English language and have a small amount of in-domain data available. Our results reveal two strategies that outperform the state-of-the-art: Translation-based methods in combination with domain-specific encoders and cross-lingual encoders plus adapters. We find that these strategies perform especially well for classifying rare phenotypes and we advise on which method to prefer in which situation. Our results show that using multilingual data overall improves clinical phenotyping models and can compensate for data sparseness.},
  url   = {https://aclanthology.org/2022.lrec-1.95}
}

@InProceedings{mcnamee-duh:2022:LREC,
  author   = {McNamee, Paul and Duh, Kevin},
  title   = {The Multilingual Microblog Translation Corpus: Improving and Evaluating Translation of User-Generated Text},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {910--918},
  abstract = {Translation of the noisy, informal language found in social media has been an understudied problem, with a principal factor being the limited availability of translation corpora in many languages. To address this need we have developed a new corpus containing over 200,000 translations of microblog posts that supports translation of thirteen languages into English. The languages are: Arabic, Chinese, Farsi, French, German, Hindi, Korean, Pashto, Portuguese, Russian, Spanish, Tagalog, and Urdu. We are releasing these data as the Multilingual Microblog Translation Corpus to support futher research in translation of informal language. We establish baselines using this new resource, and we further demonstrate the utility of the corpus by conducting experiments with fine-tuning to improve translation quality from a high performing neural machine translation (NMT) system. Fine-tuning provided substantial gains, ranging from +3.4 to +11.1 BLEU. On average, a relative gain of 21\% was observed, demonstrating the utility of the corpus.},
  url   = {https://aclanthology.org/2022.lrec-1.96}
}

@InProceedings{sato-caseli-specia:2022:LREC,
  author   = {Sato, Júlia and Caseli, Helena and Specia, Lucia},
  title   = {Multilingual and Multimodal Learning for Brazilian Portuguese},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},

  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {919--927},
  abstract = {Humans constantly deal with multimodal information, that is, data from different modalities, such as texts and images. In order for machines to process information similarly to humans, they must be able to process multimodal data and understand the joint relationship between these modalities. This paper describes the work performed on the VTLM (Visual Translation Language Modelling) framework from (Caglayan et al., 2021) to test its generalization ability for other language pairs and corpora. We use the multimodal and multilingual corpus How2 (Sanabria et al., 2018) in three parallel streams with aligned English-Portuguese-Visual information to investigate the effectiveness of the model for this new language pair and in more complex scenarios, where the sentence associated with each image is not a simple description of it. Our experiments on the Portuguese-English multimodal translation task using the How2 dataset demonstrate the efficacy of cross-lingual visual pretraining. We achieved a BLEU score of 51.8 and a METEOR score of 78.0 on the test set, outperforming the MMT baseline by about 14 BLEU and 14 METEOR. The good BLEU and METEOR values obtained for this new language pair, regarding the original English-German VTLM, establish the suitability of the model to other languages.},
  url   = {https://aclanthology.org/2022.lrec-1.97}
}

@InProceedings{jeuris-niehues:2022:LREC,
  author   = {Jeuris, Pedro and Niehues, Jan},
  title   = {LibriS2S: A German-English Speech-to-Speech Translation Corpus},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {928--935},
  abstract = {Recently, we have seen an increasing interest in the area of speech-to-text translation. This has led to astonishing improvements in this area. In contrast, the activities in the area of speech-to-speech translation is still limited, although it is essential to overcome the language barrier. We believe that one of the limiting factors is the availability of appropriate training data. We address this issue by creating LibriS2S, to our knowledge the first publicly available speech-to-speech training corpus between German and English. For this corpus, we used independently created audio for German and English leading to an unbiased pronunciation of the text in both languages. This allows the creation of a new text-to-speech and speech-to-speech translation model that directly learns to generate the speech signal based on the pronunciation of the source language. Using this created corpus, we propose Text-to-Speech models based on the example of the

recently proposed FastSpeech 2 model that integrates source language information. We do this by adapting the model to take information such as the pitch, energy or transcript from the source speech as additional input.},
  url   = {https://aclanthology.org/2022.lrec-1.98}
}

@InProceedings{macketanz-EtAl:2022:LREC,
  author   = {Macketanz, Vivien and Avramidis, Eleftherios and Burchardt, Aljoscha and Wang, He and Ai, Renlong and Manakhimova, Shushen and Strohriegel, Ursula and Möller, Sebastian and Uszkoreit, Hans},
  title   = {A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German--English Machine Translation Output},
  booktitle   = {Proceedings of the Language Resources and Evaluation Conference},
  month   = {June},
  year   = {2022},
  address   = {Marseille, France},
  publisher   = {European Language Resources Association},
  pages   = {936--947},
  abstract = {This paper presents a fine-grained test suite for the language pair German–English. The test suite is based on a number of linguistically motivated categories and phenomena and the semi-automatic evaluation is carried out with regular expressions. We describe the creation and implementation of the test suite in detail, providing a full list of all categories and phenomena. Furthermore, we present various exemplary applications of our test suite that have been implemented in the past years, like contributions to the Conference of Machine Translation, the usage of the test suite and MT outputs for quality estimation, and the expansion of the test suite to the language pair Portuguese–English. We describe how we tracked the development of the performance of various systems MT systems over the years with the help of the test suite and which categories and phenomena are prone to resulting in MT errors. For the first time, we also make a large part of our test suite publicly available to the research community.},
  url   = {https://aclanthology.org/2022.lrec-1.99}
}