Enriching Multiword Terms in Wiktionary with Pronunciation Information

Lenka Bajčetić, Thierry Declerck, Gilles Serasset

Enriching Multiword Terms in Wiktionary with Pronunciation Information

Lenka Bajčetić
Innovation Center of the School of Electrical Engineering in Belgrade
Bulevar kralja Aleksandra 73
11000 Belgrade, Serbia
lenka.bajcetic@ic.etf.ac.bg.rs

Thierry Declerck
DFKI GmbH, Multilingual Technologies
Saarland Informatics Campus D3 2
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
declerck@dfki.de

Gilles Sérasset
Université Grenoble Alpes
CNRS, Grenoble INP*, LIG
38000 Grenoble, France
gilles.serasset@imag.fr

Abstract

We report on work in progress dealing with the automated generation of pronunciation information for English multiword terms (MWTs) in Wiktionary, combining information available for their single components. We describe the issues we were encountering, the building of an evaluation dataset, and our teaming with the DBnary resource maintainer. Our approach shows potential for automatically adding morphosyntactic and semantic information to the components of such MWTs.

1 Introduction

In this paper, we describe our approach to enrich English multiword terms (MWTs) included in Wiktionary by generating pronunciation information using the existing pronunciation(s) of their sub-parts. Results of our work can also be integrated in other lexical resources, like the Open English WordNet (McCrae et al., 2020),[1] where pronunciation information has been added only for single word entries, as described in (Declerck et al., 2020a).

The main focus of our work is on generating pronunciation information for MWTs that contain (at least) one heteronym[2], as for this a specific processing of the Wiktionary data is needed, disambiguating between the different senses of the heteronym for selecting the appropriate pronunciation of this one component to be attached to the overall pronunciation. An example of such a case is given by the Wiktionary entry "acoustic bass", for which our algorithm has to specify that the pronunciation /beɪs/ (and not /bæs/) has to be selected and combined with /əˈkuːˌstɪk/. It is important to mention that although Wiktionary often lists several pronunciations for various variants of English, in this work we focus only on the standard, received pronunciation as encoded by the International Phonetic Alphabet (IPA)[3] (more about this in the Limitations Section).

Since we need to semantically disambiguate one or more components of a MWT for generating its pronunciation, our work can lead to the addition of morphosyntactic and semantic information of those components and thus enrich the overall representation of the MWTs entries, a task we have started to work on.

2 Wiktionary

Wiktionary[4] is a freely available web-based multilingual dictionary. Like other Wikimedia[5] supported initiatives, it is a collaborative project. This means that there might be inaccuracies in the resource, but the editing system is helping in mitigating this risk. The fact that Wiktionary is build by a collaborative effort means that the coverage and variety of lexical information is much larger than any single curated resource, while Wiktionary is

[1] See also https://en-word.net/

[2] The online Oxford Dictionary gives this definition: "A heteronym is one of two or more words that have the same spelling but different meanings and pronunciation, for example 'tear' meaning 'rip' and 'tear' meaning 'liquid from the eye'" https://www.oxfordlearnersdictionaries.com/definition/english/heteronym, [accessed 27.03.2023.]

[3] See https://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/

[4] https://en.wiktionary.org/

[5] https://www.wikimedia.org/

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 65–72

integrating information from expert-based dictionary resources, when their licensing conditions allow it. Nastase and Strapparava (2015) discussed already the quality (and quantity) of information included in the English Wiktionary edition, also in comparison with WordNet.[6]

Wiktionary includes, among others, a thesaurus, a rhyme guide, phrase books, language statistics and extensive appendices. Wiktionary's information also (partly) includes etymologies, pronunciations, sample quotations, synonyms, antonyms and translations.[7] Wiktionary has also developed categorization practices which classify an entry along the lines of linguistics (for example "developed terms by language") but also topical information (for example "en:Percoid fish").[8]

It has been shown that the access and use of Wiktionary can be helpful in Natural Language Processing (NLP). Kirov et al. (2016) and McCarthy et al. (2020), for example, describe work to extract and standardize the data in Wiktionary and to make it available for a range of NLP applications, while the authors focus on extracting and normalizing a huge number of inflectional paradigms across a large selection of languages. This effort contributed to the creation of the UniMorph data (http://unimorph.org/). Metheniti and Neumann (2018, 2020) describe a related approach, but making use of a combination of the HTML pages and the underlying XML dump of the English edition of Wiktionary,[9] which is covering also 4,315 other languages, but some of them with a very low number of entries.[10] Segonne et al. (2019) describe the use of Wiktionary data as a resource for word sense disambiguation tasks.

BabelNet[11] is also integrating Wiktionary data,[12] with a focus on sense information, in order to support, among others, word sense disambiguation and tasks dealing with word similarity and sense clustering (Camacho-Collados et al., 2016). The result of our work could be relevant for BabelNet, as the audio files displayed by BabelNet are not based on the reading of pronunciation alphabets but on external text-to-speech systems, which are leading to errors, as can be seen in the case of the heteronym "lead", for which BabelNet offers only one pronunciation.[13]

3 Multiword Terms in Wiktionary

Wiktionary introduces the category "English multiword terms" (MWTs), which is defined as "lemmas that are an idiomatic combination of multiple words,"[14] while Wiktionary has its page "multiword expression", categorized as a MWTs and defined as "lexeme-like unit made up of a sequence of two or more words that has properties that are not predictable from the properties of the individual words or their normal mode of combination".[15] We see these two definitions are interchangeable, since they both focus on the aspect of non-compositionality of a lexeme built from multiple words. We will therefore use in this paper the terms MWE and MWT interchangeably, but stressing that we are dealing with MWEs as they are categorized as MWTs in Wiktionary.

4 Related Work

Wiktionary is often used as a source for various text-to-speech or speech-to-text models, as described in our previous work (Bajčetić and Declerck, 2022). For instance, the work of Schlippe et al. (2010) developed a system which automatically extracts phonetic notations in IPA from Wiktionary to use for automatic speech recognition. A more recent example is the work by Peters et al. (2017) which is aimed at improving grapheme-to-phoneme conversion by utilizing

[6] See (Fellbaum, 1998) and http://wordnetweb.princeton.edu/perl/webwn for the on-line version of Princeton WordNet.

[7] See https://en.wikipedia.org/wiki/Wiktionary for more details.

[8] So that the entry "sea bass" is categorized, among others, both as an instance of "English multiword terms" and of "en:Percoid fish". The categorization system is described at https://en.wiktionary.org/wiki/Wiktionary:Categorization

[9] Wiktionary data dumps are available at https://dumps.wikimedia.org/.

[10] Details on the number of entries in the different languages contained in the English Wiktionary is given here: https://en.wiktionary.org/wiki/Special:Statistics?action=raw.

[11] See (Navigli and Ponzetto, 2010) and https://babelnet.org/.

[12] As far as we are aware of, BabelNet integrates only the English edition of Wiktionary, including all the languages covered by this edition.

[13] See the audio file associated with the two different senses of the entry for "lead": https://babelnet.org/synset?id=bn%3A00006915n&orig=lead&lang=EN and https://babelnet.org/synset?id=bn%3A00050340n&orig=lead&lang=EN.

[14] https://en.wiktionary.org/wiki/Category:English_multiword_terms. This category is an instance of the umbrella category "Multiword terms by language" see https://en.wiktionary.org/wiki/Category:Multiword_terms_by_language.

[15] https://en.wiktionary.org/wiki/multi-word_expression.

Wiktionary. Grapheme-to-phoneme is necessary for text-to-speech and automatic speech recognition systems.

Besides text-to-speech, there are various other applications which rely on extracting pronunciation information from Wiktionary. A recent tool is WikiPron (Lee et al., 2020), which is an open-source command-line tool for extracting pronunciation data from Wiktionary. It stores the extracted word/pronunciation pairs in TSV format.[16] We observe that no Wiktionary multiword terms are included in those lists. Also, no (semantic) disambiguation is provided and, for example, the word "lead" is listed twice, with the different pronunciations, but with no sense information, as WikiPron is providing solely word/pronunciation pairs. Results of our work consisting in generating pronunciation information to multiword terms could thus be included to WikiPron directly or via Wiktionary updates. In the other direction, WikiPron could be re-used for our purposes, as it harmonizes phonemic pronunciation data across various Wiktionary language editions, while the pronunciations are segmented, and stress and syllable boundary markers removed. Especially the latter is relevant for our work, as it will ease future evaluation work (see the issues described in Section 6).

Another related effort, and a very relevant resource for our approach, is DBnary.[17] DBnary extracts different types of information from Wiktionary (covering 23 languages) and represents it in a structured format, which is compliant to the guidelines of the Linguistic Linked Open Data framework.[18] In the DBnary representation of Wiktionary we find lexical entries (including words, MWEs or affixes, but without marking those explicitly, an issue that has been fixed in new release of DBnary, as this is requested for continuing our approach in the context of DBnary), their pronunciation (if available in Wiktionary), their sense(s) (definitions in Wiktionary), example sentences and DBnary glosses, which are offering a kind of "topic" for the (disambiguated) entries, but those glosses are not extracted from the category system of Wiktionary. They are taken from available information used to denote the lexical sense of the source of the translation of an entry from English to other languages.

DBnary does not include categorial information from Wiktionary, and also did not offer support for dealing with MWTs lacking pronunciation information and that contain (at least) one heteronym. Therefore, we still need(ed) to access and consult Wiktionary directly, using methods that are described in Section 5, also for building the Gold Standard for evaluating our work (MWTs in Wiktionary that are carrying pronunciation information). Hence, our results can also be integrated in DBnary, directly or via the updated Wiktionary entries. In fact, our work lead to the adaptation of DBnary, as this is briefly described in Section 5.3

5 Method

We describe in this section the various approaches we implemented and tested, leading finally to a closer cooperation with the maintainer of DBnary, as it became apparent that the release of a new version of this resource is the most efficient way for achieving and widening our goals.

5.1 Data Extraction and an Evaluation Dataset

The current version of the English edition of Wiktionary is listing 157,883 English multiword terms[19], and 75,401 expressions are categorized as "English terms with IPA pronunciation"[20]. This is quite a small number in comparison to the whole English Wiktionary, which has over 8.5 million expressions.

When we are analysing these figures, we need to be aware that they are representing the number of pages categorized as a particular category, and a Wiktionary page can often contain several lexical entries, although this is typically not the case for MWTs. Also, it is important to keep in mind that the English Wiktionary contains a lot of terms which are not English. We can see the exact number of Wiktionary pages classified as English lemmas if we look at the category itself[21]. The actual

[16] As of today, more than 3 million word/pronunciation pairs from more than 165 languages. Corresponding files are available at https://github.com/CUNY-CL/wikipron/tree/master/data.

[17] See (Sérasset and Tchechmedjiev, 2014; Sérasset, 2015) and http://kaiko.getalp.org/about-dbnary/ for the current state of development of DBnary.

[18] See (Declerck et al., 2020b) and http://www.linguistic-lod.org/.

[19] https://en.wiktionary.org/wiki/Category:English_multiword_terms [accessed 27.03.2023.]

[20] https://en.wiktionary.org/wiki/Category:English_terms_with_IPA_pronunciation [accessed 27.03.2023.]

[21] https://en.wiktionary.org/wiki/Category:English_lemmas [accessed 27.03.2023.]

number of 711,641 means that a little over 10% of English lemmas have pronunciation, while approximately 22% of all English lemmas belong in the MWT category. So there is clearly a gap that needs to be filled when it comes to pronunciation information in Wiktionary. While introducing pronunciation for the remaining 90% of lemmas seems like it has to be a manual task (or semi-automatic, using other lexical resources) - we have investigated ways to produce the missing pronunciation for numerous MWTs.

The first approach we have attempted seems to be the most straightforward, but turned out to be inefficient: download and parse the latest Wiktionary XML dump, and check for each page whether it is an English MWT using the Wiktionary API, as the corresponding category information (English multiword terms) is not included in the dump, so that it can not be accessed on the local computer. This would be simple if the size of Wiktionary dump was not so massive: more than 8.5 million entries need to be checked, which means 8.5 million requests sent to Wiktionary API. This approach was quite slow, and we thought there must be a better way for future extraction tasks that have to deal with the Wiktionary category system. Using this approach we have extracted over 98% of MWTs from Wiktionary pages and compiled a list of 153,525 multiword terms without IPA, and a gold standard of 4,979 MWTs with IPA information - we can see that only about 3% of MWTs have pronunciation information in Wiktionary.

The other approach we have followed was using the data that DBnary extracted from Wiktionary, in a structured fashion. Unfortunately, DBnary did not, at that time, encode explicitly Wiktionary MWTs. It encoded all lexical entries included in Wiktionary pages the same way, independently if they were single words, MWEs or affixes. Nevertheless, this approach was much faster, but we could only extract English multiword terms that have a blank space or a hyphen - which is not as precise as using the Wiktionary categories. We could collect 6,767 MWTs equipped with pronunciation information (in contrast to 152,082 MWTs without such information), which, combined with the data extracted with the help of the Wiktionary API, is being used as our Gold Standard for evaluating the generation of pronunciation information for MWTs.

Figure 1: The heteronymous word “bass”

We need to stress, here, that DBnary operates with lexical entries and not just pages, and therefore we had some small differences in the counted set of MWTs with pronunciations.

5.2 Generating Pronunciation Information for MWTs

As a first step, we looked at words which are unambiguous when it comes to their pronunciation. This means that a particular word has one pronunciation, even if the word has several meanings. In this case, we were not concerned with semantic ambiguity, since this is not reflected in the pronunciation, and we can easily create new pronunciation of the MWT using the pronunciations of its components. For example, “river bank” and “bank robber” both have the same sounding word “bank”, albeit its meaning is different.

But there are many words that can be included in MWTs which have pronunciation-related ambiguity. As we have previously mentioned, these words are known as heteronyms, and they have different pronunciations connected to their different meanings. Wiktionary lists over 1,000 examples of English heteronyms.[22]

In the case of MWTs that contain heteronyms, it is not straightforward to create their pronunciation by combining pronunciations of their components. Luckily, Wiktionary has other useful features, which we have exploited in this case: “Etymology” and “Derived terms” sections.

Wiktionary organizes its pages in different sections called "Etymology". We can have distinct part-of-speech (PoS) information in one Etymology section, and for each PoS different senses. Pronunciation information is distributed over the distinct Etymology sections. So that the page “bass” has 3 Etymology sections, with a total of 5 word categories. Two distinct pronunciations

[22]Listed here: https://en.wiktionary.org/wiki/Category:English_heteronyms [accessed 27.03.2023.]

are listed, whereas one pronunciation is only for the first Etymology section and the second is distributed over the other Etymology sections. We need therefore to identify the right Etymology section for extracting the correct pronunciation for the word "bass" when being a component of a MWT.

The "Derived terms" section(s) are included in the page at the level of the PoS information, and is giving us a decisive hint, as many derived terms are in fact a MWT. The MWT "black bass" is listed as a derived term of the second Etymology category of the entry "bass", and we can thus pick the associated pronunciation information for this component for building the pronunciation information for the whole MWT entry.

Using the "Etymology" and "Derived terms" sections of Wiktionary, we can make sure that we are detecting the correct lexical entry carrying the pronunciation information to produce the pronunciations of all the MWTs that contain it, as a first manual comparison with our evaluation dataset confirms.

In this context, we discovered an even easier approach, which is still to be implemented: if in the list of "Derived terms" we find one MWT with pronunciation information, we can segment this pronunciation information and propagate it to all other MWTs containing the one word of which the MWT is listed as a "Derived term". This approach is currently under evaluation, and seems to be more accurate, as in the "Derived terms" section only one pronunciation type is given, while in the entries of the single words, there are different types of pronunciation information.

To summarize: The access to the "Derived terms", coupled with the "Etymology" classification, is the key that allows us not only to compute the pronunciation information, but also add morphosyntactic and semantic information to the components of a MWT.

5.3 A new Release of DBnary

As we already mentioned, DBnary was not explicitly marking MWEs in its data extracted from Wiktionary. DBnary was also not considering the "Derived terms" sections. The maintainer of DBnary could offer this information in a new update, and therefore we focus in the current and future work on the use of DBnary for achieving our goals.

An additional aspect that motivated our decision is the fact that DBnary is exclusively making use

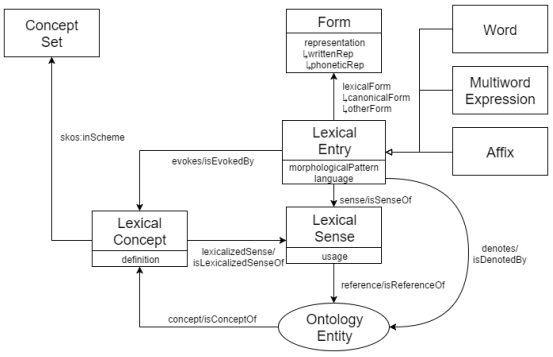

Figure 2: The core module OntoLex-Lemon. Taken from https://www.w3.org/2016/05/ontolex/#core

of accepted specifications and standards for representing its data. Lexical data in DBnary is represented using the Linked Open Data (LOD) principles[23] and as such it is using RDF[24] as its representation model. It is freely available and may be either downloaded or directly queried on the internet. DBnary uses the *ontolex* standard vocabulary (Cimiano et al., 2016),[25] displayed in Figure 2, to represent the lexical entries structures, along with other widely accepted RDF-based vocabularies in the field of language technologies.

As DBnary is making use of the OntoLex-Lemon model, we can take advantage of the existence of the "Decomposition" module of this model.[26] We display in Figure 3 the graphical representation of this module.

We can directly map the data extracted from the "Derived terms" sections in Wiktionary to elements of the Decomposition module of Ontolex, and mark the full lexical description of a single word as a "ontolex:subterm" of a MWE encoded in the Ontolex model.

As a result, the recent adaptations of DBnary allow not only to generate pronunciation informa-

[23] See https://www.w3.org/wiki/LinkedData for more information on those principles

[24] The Resource Description Framework (RDF) model is a graph based model for the representation of data and metadata, using URIs to represent resources (nodes) and properties (edges).

[25] See also the specification document at https://www.w3.org/2016/05/ontolex/.

[26] The specification of OntoLex-Lemon describes "Decomposition" in those terms: "Decomposition is the process of indicating which elements constitute a multiword or compound lexical entry. The simplest way to do this is by means of the subterm property, which indicates that a lexical entry is a part of another entry. This property allows us to specify which lexical entries a certain compound lexical entry is composed of.". Taken from https://www.w3.org/2016/05/ontolex/#decomposition-decomp

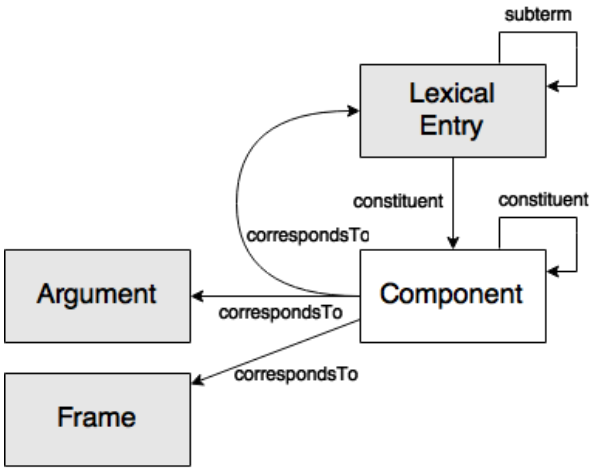

Figure 3: The Decomposition module of OntoLex-Lemon. Taken from https://www.w3.org/2016/05/ontolex/#decomposition-decomp

tion for MWTs contained in the English edition of Wiktionary, but also to add all the lexical information encoded in the lexical description of the components of such MWTs, and to represent this information in such a way that the new data set can be published on the Linguistic Linked Open Data cloud.

6 An initial Evaluation Study

In order to evaluate the newly created pronunciations, we use those MWTs which already carry pronunciation information in Wiktionary. In a first "naive" approach, we just compared the result of combining the extracted pronunciation information from the components of the MWTs with those MWTs which are equipped with pronunciation in Wiktionary. This simple string matching lead to poor results, as it might have been expected. One of the reasons being that in some cases the pronunciation information included in the MWT is containing either space(s), suprasegmental information, or other markers. The combination of pronunciation information extracted from the components do not contain those additional information (at least not in the same way).

Another issue we were confronted with, lies in the fact that in many cases, Wiktionary is listing more than one pronunciation information for a single word. Our algorithm needs to be tuned in order to select only the one pronunciation information that is included in the corresponding MWT.

Some editing of the evaluation set is also needed, towards the creation of an evaluation set that is containing no suprasegmental pronunciation information (and other markers) or spaces. A first analysis of such a cleaned evaluation data set showed already an improved computation of recall and precision. We plan to use for this also the data set generated by the WikiPron initiative (see the description in Section 4.

7 Conclusion and Future Work

We described work in progress consisting in adding automatically generated pronunciation information to MWTs included in the English edition of Wiktionary. The current outputs of our work consist of an evaluation data set for this task, and a set of algorithms for accessing specific information in Wiktionary. We motivated our decision for teaming with the DBnary maintainer, as we can this way widen our goals to the inclusion of morphosyntactic and semantic information to the components of MWTs included in Wiktionary.

Future work includes adding the pronunciations to Wiktionary and enriching other lexical resources, beginning with the Open English WordNet. We will also extend our work to the other language editions of Wiktionary covered by DBnary, at least dealing with the addition of morphosyntactic and semantic information to the components of MSTs, in those languages.

Limitations

While our approach can probably be transferred to other languages, in cases where the Wiktionary structure for those languages is similar, there is one aspect of pronunciation extraction and combination that we have not discussed and this concerns the pronunciation(s) of variants of English, which are included in Wiktionary, like British, General American, Irish, Canadian, Australian and New Zealand English. In our current work we have decided to focus on the non-specific variant, so for now we "overlook" some pronunciation(s) of entries, as we did not want to mix different variants and produce potentially unusable new pronunciations. The standard version is typically considered to be "Received pronunciation", commonly known as "BBC English".[27] However, we would want to include all these variants in our future work. The approach would follow the same

[27] https://en.wiktionary.org/wiki/Received_Pronunciation

principle as explained in the paper, with one extra layer of variant matching.

Another limitation of our work lies in the fact that Wiktionary is ever-changing. So anything done at one point in time needs to be re-done in the future due to changes in the data and also newly added data. The fact that Wiktionary grows quite fast means that the best approach would be incremental or recursive in some way, and automatically check for newly added pronunciations which can create new MWEs pronunciations, while also confirming that the previously created ones have not been altered and need updating. This is a reason why we teamed with the makers of DBnary for this, as DBnary is updated twice a month.

Ethics Statement

We consider our work to have a broad impact because Wiktionary is widely used across the world, and it is free and open-source. Additionally, we plan to include the output of our research into the Open English WordNet and other lexical resources, which are free to use and open-source. We hope that in this way the result of our work can potentially be useful to people all around the world who read or speak English, as well as text-to-speech (and possibly speech-to-text) systems which are gaining popularity and are very important for the visually impaired community, among others.

We do not see any ethical issue related to the generation of additional information to be attached to Wiktionary MWTs and their components.

Acknowledgements

The presented work is pursued in the context of the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), 731015). The DFKI contribution is also pursued within the LT-BRIDGE project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194.

References

Lenka Bajčetić and Thierry Declerck. 2022. Using Wiktionary to Create Specialized Lexical Resources and Datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3457–3460, Marseille, France. European Language Resources Association.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artif. Intell.*, 240:36–64.

Philipp Cimiano, John McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report, 10 May 2016. Technical report, W3C.

Thierry Declerck, Lenka Bajcetic, and Melanie Siegel. 2020a. Adding pronunciation information to wordnets. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 39–44, Marseille, France. The European Language Resources Association (ELRA).

Thierry Declerck, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Saurí, Deirdre Lee, Stefania Racioppa, Jamal Abdul Nasir, Matthias Orlikowsk, Marta Lanau-Coronas, Christian Fäth, Mariano Rico, Mohammad Fazleh Elahi, Maria Khvalchik, Meritxell Gonzalez, and Katharine Cooney. 2020b. Recent Developments for the Linguistic Linked Open Data Infrastructure. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5660–5667, Marseille, France. European Language Resources Association.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).

Eleni Metheniti and Günter Neumann. 2018. Wikinflection: Massive semi-supervised generation of multilingual inflectional corpus from Wiktionary. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings universitet.

Eleni Metheniti and Günter Neumann. 2020. Wikinflection corpus: A (better) multilingual, morpheme-annotated inflectional corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. LREC.

Vivi Nastase and Carlo Strapparava. 2015. knoWitiary: A Machine Readable Incarnation of Wiktionary. *Int. J. Comput. Linguistics Appl.*, 6:61–82.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. *CoRR*, abs/1708.01464.

Tim Schlippe, Sebastian Ochs, and Tanja Schultz. 2010. Wiktionary as a source for automatic pronunciation extraction. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 2290–2293. ISCA.

Vincent Segonne, Marie Candito, and Benoît Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Gothenburg, Sweden. Association for Computational Linguistics.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6:355–361.

Gilles Sérasset and Andon Tchechmedjiev. 2014. Dbnary : Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, Paris, France.