Adding Information to Multiword Terms in Wiktionary

Thierry Declerck[1], **Lenka Bajčetić**[2], **Gilles Sérasset**[3]

[1] DFKI GmbH, Multilingual Technologies, Saarland Informatics Campus D3-2, D-66123 Saarbrücken, Germany

[2] Innovation Center of the School of Electrical Engineering in Belgrade, Bulevar kralja Aleksandra 73,
11000 Belgrade, Serbia

[3] Université Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France

E-mail: declerck@dfki.de, lenka.bajcetic@ic.etf.ac.bg.rs, gilles.serasset@imag.fr

Abstract

We describe ongoing work dealing with the potential "auto-enrichment" of "Multiword terms" (MWTs) that are included in the English edition of Wiktionary. The idea is to use and combine information contained in the lexical components of the MWTs and to propagate this extracted and filtered information into the lexical description of the MWTs, as those are typically equipped with less lexical information as it is the case for their lexical components. We started our work with the generation of pronunciation information for such MWTs, on the base of the pronunciation information available for their components. We present in this paper first achievements but also issues we encountered. Addressing those issues lead us to consider additional resources for supporting our approach, like DBnary and WikiPron. This step was ultimately leading to suggestions of adaptations for those additional resources, which, in the case of DBnary, are already implemented. We are currently extending our approach to a morphosyntactic and semantic enrichment of the English MWTs in Wiktionary.

Keywords: Multiword terms; Wiktionary; lexical enrichment; linguistic linked data

1. Introduction

We describe an approach aiming at enriching English multiword terms (MWTs) included in Wiktionary by generating lexical information gained by using, filtering and combining available lexical descriptions of their lexical components.

We started our work with the generation of pronunciation information, as we noticed that a vast majority of English MWTs in Wiktionary are lacking this type of information. While designing a potential evaluation dataset for the pronunciations generated by our approach, we noticed that only around 3% of MWTs are carrying pronunciation information. We also discovered that other complex lexical constructions (affix + word, or word + affix) are often lacking pronunciation information. We collected for the evaluation dataset 6,768 MWT entries with pronunciation (compared with 252,082 MWT entries that are lacking such information). Our approach for generating pronunciation information for MWTs consisted in combining the pronunciation information included in the lexical description of their components. Results of this work can be integrated in other lexical resources, like the Open English WordNet (McCrae et al., 2020),[1] where pronunciation information has been added for now only for single word entries, as described in (Declerck et al., 2020a).

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

A specific issue emerged for the generation of pronunciation information for MWTs that contain (at least) one heteronym.[2] For this type of lexical entry a specific processing is needed, disambiguating between the different senses of the heteronym for extracting the appropriate pronunciation of this one lexical component to be selected to form the overall pronunciation of the MWT. 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 /əˈkuː.stɪk/. It is important to mention that Wiktionary often lists several pronunciations for various variants of English. In this work we focus on the standard, received pronunciation for English, as encoded by the International Phonetic Alphabet (IPA).[3]

Since there are cases for which we need to semantically disambiguate one or more lexical components of a MWT for generating its pronunciation, our work can also lead to the addition of disambiguated morphosyntactic and semantic information of those components to the lexical description of MWTs, and thus enrich the overall representation of the MWTs entries beyond the generation of pronunciation information. This is a task we have started to work on.

In this paper, we describe first briefly the way multiword terms (MWTs) are introduced in Wiktionary. We summarize then the various approaches we followed for both designing an evaluation dataset and generating pronunciation information, dealing for now with the English edition of Wiktionary. We discuss issues we encountered, and which lead to the consultation of related resources, like DBnary (Sérasset & Tchechmedjiev, 2014; Sérasset, 2015) and WikiPron (Lee et al., 2020). While the cooperation with DBnary has been already established and resulted in improvements of our approach and an adaptation of DBnary itself, which we describe in some details, we are starting with the formulation of suggestions for adaptation for WikiPron. We present our first step towards the enrichment of MWTs with morphosyntactic and semantic information extracted from their components. We close the paper with conclusive remarks and presenting future work.

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 coverage and variety of lexical information is much larger than any single curated resource, while Wiktionary is integrating information from expert-based dictionary resources, when their licensing conditions allow it. Nastase and Strapparava (2015) gave some details on the quality (and quantity) of information included in the English Wiktionary edition, also in comparison with WordNet.[6]

[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 20.04.2023.]

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

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

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

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

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 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"). So that the entry "sea bass" is categorized, among others, both as an instance of "English multiword terms" and of "en:Percoid fish".[8]

3. Multiword Terms in Wiktionary

The version of the English edition of Wiktionary we worked with is listing 159,169 English multiword terms,[9] and 75,646 expressions are categorized as "English terms with IPA pronunciation".[10] This is quite a small number in comparison to the whole English Wiktionary, which has over 8,633,770 pages (among those, 7,387,538 are classified as content pages[11]). When we analyse 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.[12] The actual number of 714,732 means that a little over 10% of English lemmas have pronunciation, while approximately 22% of all English lemmas belong in the MWT category. There is clearly a gap that needs to be filled when it comes to pronunciation information in Wiktionary. While introducing pronunciation for the remaining (non MWTs) 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.

4. A first Approach

We designed a computer program to extract from the Wiktionary XML dumps[13] the pronunciation information from the component words and to combine them for the corresponding MWT, limiting our work to MWTs with two component words, which are building a majority of the cases, and which are well described in Wiktionary, with clear links to pages containing their component parts, while MWTs having more components are more poorly represented in Wiktionary.

This way, we can straightforwardly create a huge amount of pronunciation information that we can add to English MWTs included in Wiktionary. However, there is this issue concerning the cases in which a MWT is containing at least one heteronym. As the Wiktionary entry of the MWT is pointing back for its lexical components to Wiktionary pages (which often contain more than one lexical entry), but not to the specific entry

[7] See https://en.wikipedia.org/wiki/Wiktionary for more details.
[8] The categorization system is described at https://en.wiktionary.org/wiki/Wiktionary:Categorization
[9] https://en.wiktionary.org/wiki/Category:English_multiword_terms [accessed 20.04.2023.]
[10] https://en.wiktionary.org/wiki/Category:English_terms_with_IPA_pronunciation [accessed 20.04.2023.]
[11] See https://en.wiktionary.org/wiki/Special:Statistics [accessed 20.04.2023]
[12] https://en.wiktionary.org/wiki/Category:English_lemmas [accessed 20.04.2023.]
[13] See https://dumps.wikimedia.org/ for more details

with the specific sense, we needed to adapt our approach and go into a deeper parsing of Wiktionary, adding complexity to our program. This point let us consider the use of already existing tools that are extracting information from Wiktionary, two of those tools - DBnary and WikiPron - being described in Section 1.2, and in Sections 6 and 7 respectively.

This was particularly relevant for the design of an evaluation dataset, as for this we had to query the category system of Wiktionary, which is not included in the available XML dumps. We had thus to make use of the Wiktionary API, which is a RESTful interface that allows programmers to access the data contained in the Wiktionary dictionary through standard HTTP requests. It may be used to query for definitions, translations, links or categories of a specific Wiktionary page. In our cases, we planned to use it to query each page for its categories. This would be an easy way to go if the size of English edition of Wiktionary was not so massive: more than 8.6 million entries need to be checked, which means 8.6 million requests sent to Wiktionary API. This is quite slow and if not done correctly will lead to being blacklisted from the Wiktionary website. Using this approach, we have extracted over 98% of MWTs from Wiktionary and compiled a list of 153,525 multiword terms without IPA, and a gold standard of 4,979 MWTs with IPA - we can see that only about 3% of MWTs have pronunciation information in Wiktionary. However, this approach was very time-consuming and can only be applied to a specific version of Wiktionary. Hence, as the Wiktionary data is always growing, new MWTs introduced in Wiktionary will not benefit from this work. This is the reason why we tried to reproduce our experiment using the DBnary dataset, which is regularly updated. The move to DBnary offered us some more MWTs with IPA pronunciation included in Wiktionary, resulting in the (current) total number of 6,768 MWT entries with pronunciation.

This work was needed in order to build an evaluation dataset. We aim at an "internal" evaluation of our approach, as a number of MWTs in Wiktionary are in fact equipped with pronunciation information, like "sea bass" (in the IPA encoding /ˈsiːbæs/), so that we can compare our pronunciation extraction applied to "sea" and "bass" and see if it yields the correct pronunciation from the heteronym "bass". We encountered in this context a number of Wiktionary-related issues . One issue being, that in some cases suprasegmental information is encoded in the IPA transcription of either the component(s) or in the IPA transcription associated with the MWT, so that a proper string matching approach can not be implemented. Another issue being that sometimes syllable boundaries are marked, and sometimes not. And in some cases, the IPA transcription associated with the MWT in Wiktionary is just concatenating the two IPA codes, while in other cases, a blank is introduced between the two IPA codes. We have also some issues related to the regional encodings, as sometimes we have only the US IPA code or the UK IPA code. Last but not least, sometimes two alternative IPA transcriptions are given for a single word entry, while only one is present in the IPA transcription of the corresponding MWT entry. Those issues also lead us to consider for the building of the evaluation dataset the use of the WikiPron resource, which is described in Sections 1.2 and 7.

5. Related Work

Wiktionary is often used as a source for various text-to-speech or speech-to-text models. 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 Wiktionary. Grapheme-to-phoneme is necessary for text-to-speech and automatic speech recognition systems.

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.[14] 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, while taking into consideration heteronyms, could thus be included in WikiPron directly or via Wiktionary updates. But in its actual form, WikiPron can 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 can be on request removed. Especially the latter is relevant for our work, as it will ease future evaluation work (see the issues described in Section 4). This dataset and its relevance for our work, while also discussing some shortcomings, are described in more details in Section 7.

BabelNet (Navigli & Ponzetto, 2010)[15] is one of the resources that is integrating Wiktionary data,[16] 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.[17]

A very relevant resource for our approach is DBnary (Sérasset & Tchechmedjiev, 2014; Sérasset, 2015).[18] 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.[19] In the DBnary representation of Wiktionary we find lexical entries (including words, multi word expressions (MWEs) or affixes, but without marking those sub-classes of lexical entries 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.

[14] 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.

[15] See also https://babelnet.org/.

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

[17] 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.

[18] See http://kaiko.getalp.org/about-dbnary/ for the current state of development of DBnary.

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

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 4, also for designing the dataset 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 6.

6. Cooperation with DBnary

DBnary is representing the lexical information extracted from Wiktionary using the Linked Open Data (LOD) principles[20] and as such it is using RDF[21] as its representation model. It is freely available and may be either downloaded or directly queried on the internet. DBnary uses the *OntoLex-Lemon* standard vocabulary (Cimiano et al., 2016),[22] displayed in Figure 1 to represent the lexical entries structures, along with *lexvo* (de Melo, 2015) to uniquely identify languages, *lexinfo* (Cimiano et al., 2011)[23] and *Olia* (Chiarcos & Sukhareva, 2015)[24] for linguistic data categories.

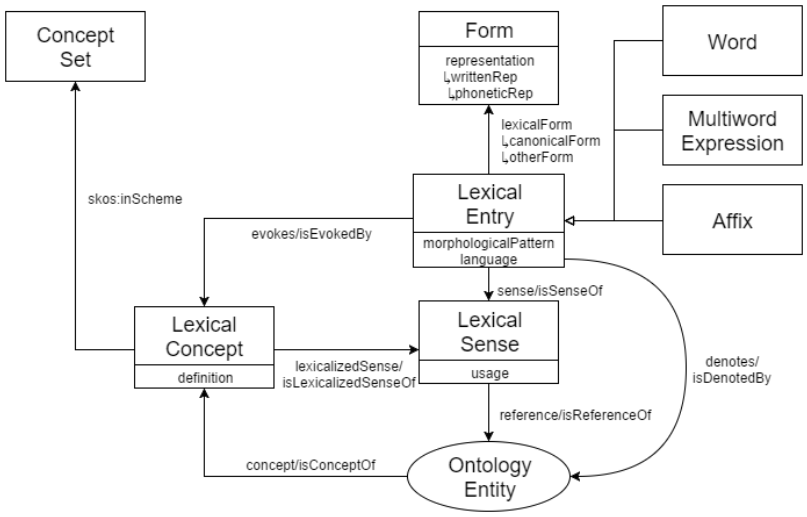

Figure 1: The core module OntoLex-Lemon. Taken from https://www.w3.org/2016/05/ontolex/#core

While trying to reproduce with the DBnary engine the work briefly described in Section 4, we noticed that DBnary was lacking some information. First, Wiktionary multiword terms were not marked explicitly. Second, derivation relations between single word lexical entries and MWTs, in which they occur, were not extracted, while this information is

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

[21] 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).

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

[23] The latest version of the lexinfo ontology can be downloaded at https://lexinfo.net/.

[24] The "Ontologies of Linguistic Annotation (OLiA)" is available at https://acoli-repo.github.io/olia/.

crucial for the disambiguation of components of MWTs that are heteronyms. The DBnary maintainer[25] tuned the extraction program to fix these identified lacks.

These missing elements were added and are now available in versions starting from February 2023. The extraction program now correctly *types* English Wiktionary entries either as `ontolex:Word` or as `ontolex:MultiWordExpression` (for the MWTs). Moreover, derivation relations are now extracted and available in the graph using `dbnary:derivesFrom` transitive property.

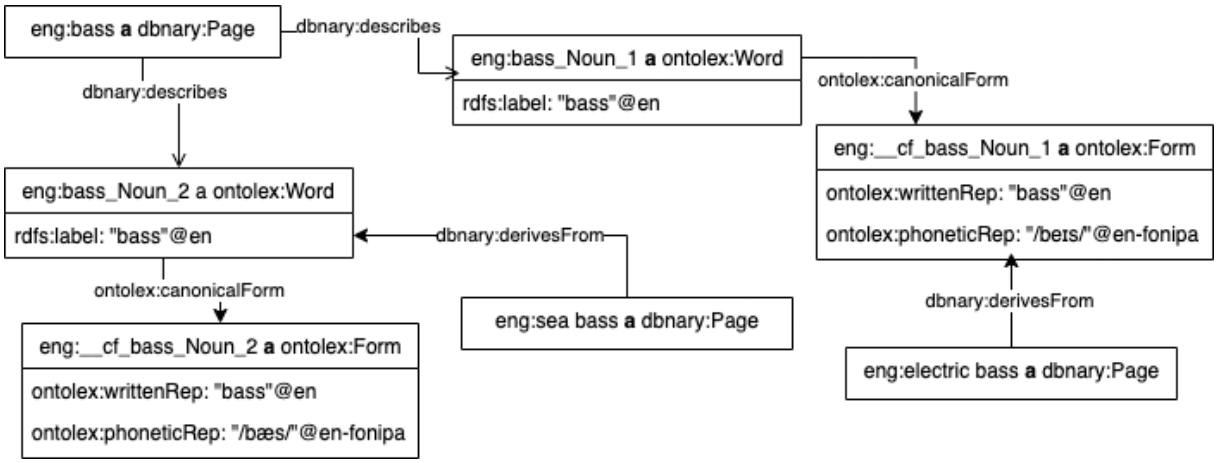

Figure 2: A very small extract of the DBnary graph showing the DBnary page *bass* and 2 of the lexical entries it describes (bass_Noun_1 [sound, music, instrument] and bass_Noun_2 [perch, fish]) and their respective canonical forms. The pages *sea bass* and *electric bass* are also represented with their derivation relations.

Figure 2 shows an example of the organisation of two heteronym lexical entries described by the same page, along with their canonical forms (with written and phonetic representations). Figure 2 also shows how the derivation relation is modelled in DBnary, using the transitive `dbnary:derivesFrom` property. It must be noted that in Wiktionary original data, the derivation links point to Wiktionary pages but not to Wiktionary entries, hence, the DBnary modelling reflects this as it is usually difficult to automatically choose which lexical entry (or entries) is (are) the valid target of the derivation relation. But, applying the property in the inverse direction (could be named `dbnary:derivesTo`), the subject/source of the relation is a lexical entry within a Wiktionary page, pointing to a MWT page. As MWT pages consist mainly of only one lexical entry, we can precisely establish a "subterm" relation between a single lexical entry and the MWTs it occurs in, combining if needed both "directions" of use of the property. This point is very important, as it allows projecting all the lexical information of the single lexical entry to the component it builds within a MWT, as this is briefly presented in Section 8.

[25] The DBnary extraction programs are open source and available at: https://gitlab.com/gilles.serasset/dbnary/ where issues can be added to ask for correction or enhancement of the extractors. It is also possible to fix the extractors and create a Merge Request.

7. About Wikipron

We were also confronted with issues with the pronunciation information in various language editions of Wiktionary, as sometimes suprasegmental information or syllables boundaries are present and sometimes not, or the fact that sometimes we have only the phonetic IPA transcription, sometimes only the phonemic transcriptions and sometimes both associated with a Wiktionary page and their entries. Those issues are building an obstacle for the creation of a clean evaluation dataset. Searching for help for this, we looked in more details at the WikiPron resource, as it is providing for a differentiated analysis of the extracted pronunciation information from Wiktionary. WikiPron is also proposing a cleaning of certain pronunciation information. The WikiPron data set is being used for example in an investigation on what phonological information is encoded in character embeddings (Boldsen et al., 2022). But contrary to the authors of this study, we would not call Wikipron a "dictionary", as we discovered certain issues, that would need to be addressed if the resource should be called a "dictionary", in a lexicographic sense.

A first issue (already discussed above) is the fact that WikiPron does not consider the extraction of pronunciation information associated with Wiktionary MWTs – although we think that the tool could (and should) extract and deliver the word-IPA pairs for those MWTs. But, as in the case of DBnary, this should be an easy "fix" to implement.

A second issue, more significant, is the fact that entries that have more than one IPA transcription are encoded in the word-IPA codes pairs as two different units. So for example, for UK English:

electric ə ˈl ɛ k t ɹ ɪ k
electric ɪ ˈl ɛ k t ɹ ɪ k

This can give the impression that we are dealing with 2 different lexical entries, as WikiPron represents in the same way the two different pronunciations for "lead", which is a heteronym and which should thus be considered as having two different lexical entries with different pronunciations **and** meanings:

lead l ɛ d
lead l iː d
lead l i d

whereas the two last pronunciations are variant for the second meaning (in fact, the last pronunciation corresponds to a misspelling of the verb).[26] A better TSV representation for both words would be:

electric ə ˈl ɛ k t ɹ ɪ k | ɪ ˈl ɛ k t ɹ ɪ k

lead l ɛ d
lead l iː d | l i d

We note that this way of presenting those cases of pronunciation information can be easily represented in OntoLex-Lemon, and could therefore be encoded directly in DBnary, contributing to another adaptation of this linked data compliant resource.

[26] See https://en.wiktionary.org/wiki/lead#Etymology_3

8. Extending the Approach to the Addition of morphosyntactic and semantic Information to MWTs

In addition to pronunciation creation and enrichment, our work can lead to another improved description of Wiktionary multiword terms (represented in DBnary as instances of the class `ontolex:MultiWordExpression`), as we can (in a next step) also add the disambiguated morphosyntactic and semantic information associated to hypernyms included in MWTs, taking as a departure point the senses used in Wiktionary itself.

As DBnary is making use of the OntoLex-Lemon model, we can take advantage of the availability of its "Decomposition" module,[27] which is graphically displayed in Figure 3.

We can observe that the property `decomp:subterm` of the Decomposition module is equivalent to the property `dbnary:derivesFrom`, recently introduced in DBnary, in order to represent the Wiktionary section "Derived terms" (see Figure 2 for comparison). Therefore, we can just map the `rdf:Object` of `dbnary:derivesFrom` to the `rdf:Object` of `decomp:subterm`, while the `rdf:Subject` of `decomp:subterm` is the MWT itself, as can be seen in Listing 20.1.

As a result, the recent adaptations of DBnary allow not only to generate pronunciation information for MWTs contained in the English edition of Wiktionary, but also to add morphosyntactic and semantic information to the components of such MWTs, and to encode this information in such a way that the new data set can be published on the Linguistic Linked Open Data cloud.

```
1 :electric_bass_lex a
2       ontolex:MultiwordExpression ;
3   decomp:subterm eng:electric_Adjective_1 ;
4   decomp:subterm :eng:bass_Noun_1 .
```

Listing 20.1: The (simplified) representation of "electric bass" using the Decomposition module of OntoLex-Lemon, with links to lexical data encoded in DBnary

Using this module, we can explicitly encode the morphosyntactic, semantic and domain information of the components of MWTs, which are only implicitly present in Wiktionary. For our example, we know yet that "electric" has PoS "adjective" (Wiktionary lists also a nominal use of the word) and "bass" the PoS "noun" (Wiktionary lists also adjectival and verbal uses), while semantically disambiguating the components of the MWT (in the full DBnary representation, the "ontolex:Word": "eng:bass_Noun_1" is linked to the corresponding instances of "ontolex:Sense". And in fact, we can then link to a corresponding Wikidata entry for "bass guitar" (https://www.wikidata.org/wiki/Q46185) and the one for "electricity" (https://www.wikidata.org/wiki/Q12725)

9. Conclusion and future Work

We described in this paper ongoing work on computing lexical information for multiword terms (MWTs) included in Wiktionary. While progressing, we were repeatedly confronted

[27] 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

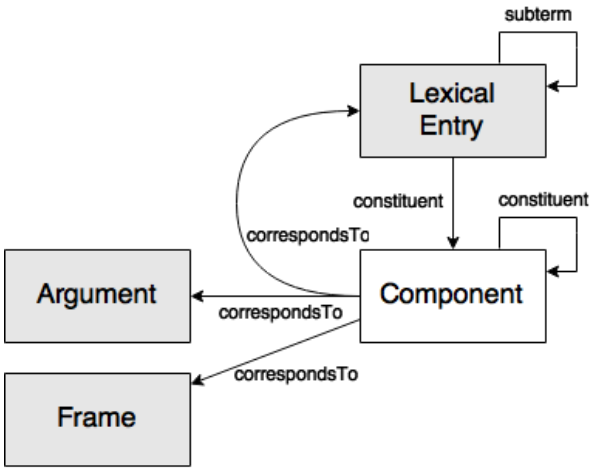

Figure 3: The Decomposition module of OntoLex-Lemon. Taken from https://www.w3.org/2016/05/ontolex/#decomposition-decomp

with issues, and we therefore investigated the combined use of other resources resulting from the extraction of information from Wiktionary. We got this way acquainted with the DBnary resource, which is offering a Linked Open Data compliant representation of lexical information extracted from Wiktionary, using at its core the OntoLex-Lemon model and other Semantic Web based vocabularies. As it was immediately clear that using the extraction engine of DBnary is massively easing our work, we teamed with the maintainer of DBnary, who adapted the extraction engine for our needs. Thanks to this cooperation, we discovered also that we can not only generate pronunciation information for MWTs, but that we can also in a straightforward manner extract morphosyntactic and semantic information from the components of MWTs and add those to the lexical description of the MWTs. The enriched information can be encoded in a principled way in OntoLex-Lemon. This will lead to the generation of a new dataset for English MWTs within the Linguistic Linked Data framework. As a result, the DBnary engine is now more than an extractor from Wiktionary and a mapper to an LOD compliant representation, as it generates lexical information that can be used for enriching existing lexical resources.

While confronted with issues related to the precise IPA encoding of pronunciation in Wiktionary, we got acquainted with the WikiPron resource, which is helping us for the building of an evaluation dataset for our pronunciation generation to be associated with MWTs. We also discovered some issues with WikiPron that would need to be addressed, as we want to add elements of this very relevant resource in a lexical framework.

Both DBnary and WikiPron are tools and resources with a large multilingual coverage, a fact that will help us to extend our work to other languages than English.

10. 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 in the context of the LT-BRIDGE project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194.

We would like to thank the anonymous reviewers for their helpful comments.

11. References

Boldsen, S., Agirrezabal, M. & Hollenstein, N. (2022). Interpreting Character Embeddings With Perceptual Representations: The Case of Shape, Sound, and Color. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6819–6836. URL https://aclanthology.org/2022.acl-long.470.

Camacho-Collados, J., Pilehvar, M.T. & Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artif. Intell.*, 240, pp. 36–64. URL https://doi.org/10.1016/j.artint.2016.07.005.

Chiarcos, C. & Sukhareva, M. (2015). OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4), pp. 379–386. Publisher: IOS Press.

Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1), pp. 29–51. URL https://www.sciencedirect.com/science/article/pii/S1570826810000892.

Cimiano, P., McCrae, J. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report, 10 May 2016. Technical report, W3C. URL https://www.w3.org/2016/05/ontolex/.

de Melo, G. (2015). Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*, 6(4), pp. 393–400.

Declerck, T., Bajcetic, L. & Siegel, M. (2020a). Adding Pronunciation Information to Wordnets. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. Marseille, France: The European Language Resources Association (ELRA), pp. 39–44. URL https://aclanthology.org/2020.mmw-1.7.

Declerck, T., McCrae, J.P., Hartung, M., Gracia, J., Chiarcos, C., Montiel-Ponsoda, E., Cimiano, P., Revenko, A., Saurí, R., Lee, D., Racioppa, S., Abdul Nasir, J., Orlikowsk, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M.F., Khvalchik, M., Gonzalez, M. & Cooney, K. (2020b). Recent Developments for the Linguistic Linked Open Data Infrastructure. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 5660–5667. URL https://aclanthology.org/2020.lrec-1.695.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA: MIT Press.

Lee, J.L., Ashby, L.F., Garza, M.E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A.D. & Gorman, K. (2020). Massively Multilingual Pronunciation Modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4223–4228. URL https://www.aclweb.org/anthology/2020.lrec-1.521.

McCrae, J.P., Rademaker, A., Rudnicka, E. & Bond, F. (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)*. Marseille, France: The European Language Resources Association (ELRA), pp. 14–19. URL https://aclanthology.org/2020.mmw-1.3.

Nastase, V. & Strapparava, C. (2015). knoWitiary: A Machine Readable Incarnation of Wiktionary. *Int. J. Comput. Linguistics Appl.*, 6, pp. 61–82.

Navigli, R. & Ponzetto, S.P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for*

Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, pp. 216–225. URL https://aclanthology.org/P10-1023.

Peters, B., Dehdari, J. & van Genabith, J. (2017). Massively Multilingual Neural Grapheme-to-Phoneme Conversion. *CoRR*, abs/1708.01464. URL http://arxiv.org/abs/1708.01464. 1708.01464.

Schlippe, T., Ochs, S. & Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In T. Kobayashi, K. Hirose & S. Nakamura (eds.) *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, pp. 2290–2293. URL http://www.isca-speech.org/archive/interspeech_2010/i10_2290.html.

Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web*, 6, pp. 355–361.

Sérasset, G. & Tchechmedjiev, A. (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*. Reyjkjavik, Iceland, p. to appear. URL http://hal.archives-ouvertes.fr/hal-00990876.