# Towards the Addition of Pronunciation Information to Lexical Semantic Resources

**Thierry Declerck**
German Research Center for AI
Multilinguality and Language Technology
Stuhsatzenhausweg 3
D-66123 Saarbrücken Germany
declerck@dfki.de

**Lenka Bajčetić**
Austrian Centre for Digital Humanities and
Cultural Heritage
Sonnenfelsgasse 19
Wien 1010, Austria
lenka.bajcetic@oeaw.ac.at

## Abstract

This paper describes ongoing work aiming at adding pronunciation information to lexical semantic resources, with a focus on open wordnets. Our goal is not only to add a new modality to those semantic networks, but also to mark heteronyms listed in them with the pronunciation information associated with their different meanings. This work could contribute in the longer term to the disambiguation of multi-modal resources, which are combining text and speech.

## 1 Introduction

The work described in this paper aims at enriching lexical semantic databases by adding the modality of pronunciation, primarily targeting in our current work the Open English WordNet (McCrae et al., 2019a, 2020).[1] Pronunciation information is typically not associated with WordNet, but can be particularly relevant within the vision of contributing directly or indirectly to integrated lexical resources and architectures, like the ELEXIS Dictionary Matrix (McCrae et al., 2019b) or BabelNet (Navigli and Ponzetto, 2010), as well as text-to-speech systems which use WordNet or WordNet-based lexical resources or tools.

In a number of cases, homographs with different meanings are also characterised by different pronunciations. This can be the case across syntactic categories, but also within one category, like for example for the noun “lead”,[2] which is having a different pronunciation per sense, as this is exemplified in the combination of the IPA[3] code [/lɛd/] and the definition:

> (“A heavy, pliable, inelastic metal element, having a bright, bluish color, but easily tarnished; both malleable and ductile,though with little tenacity. It is easily fusible, forms alloys with other metals, and is an ingredient of solder and type metal. Atomic number 82, symbol Pb (from Latin plumbum).”)

and of the IPA code [/liːd/] and the definition:

> (“The act of leading or conducting; guidance; direction, course”).

This phenomenon is called “heteronymy”. Although they share the same spelling, heteronyms have two different possible pronunciations that are associated with two (or more) different meanings (Martin et al., 1981). By definition, these words are homographs which are not homophones. They can be considered as the opposite of polyphones, which are words with different pronunciations that are not associated with different meanings. Typical heteronym examples in English include “tear” , “bow”, and “row”.

The frequency of heteronymy varies across different languages. For example, as for today, Wiktionary counts 723 cases for English,[4] while only 21 cases are listed for French.[5] But the number of concerned entries increases considerably if we take into account all the derived terms (including

[1] See also https://github.com/globalwordnet/english-wordnet.

[2] The two pronunciation and definition pairs for the noun “lead” displayed here are taken from the XML dump of the English edition of Wiktionary. The human readable page can be consulted at https://en.wiktionary.org/wiki/lead#Noun.

[3] IPA stands for ”International Phonetic Alphabet”. See also https://www.internationalphoneticassociation.org/.

[4] https://en.wiktionary.org/wiki/Category:English_heteronyms, [consulted: 2021.01.28]

[5] https://en.wiktionary.org/wiki/Category:French_heteronyms, [consulted: 2021.01.28]

compounds and phrasal expressions) in which a heteronym entry is occurring. So, for the "metal" sense of the "lead" entry, Wiktionary is listing 77 derived terms, 32 of them being currently included as an entry in the dictionary. Some of them are carrying pronunciation information ("leadsman"), and some are not ("lead pencil"). Similarly, for the "curved" sense of "bow" Wiktionary lists 19 derived terms, like for example "longbow", all included as an entry in the dictionary. Some of them are also not carrying pronunciation information, like for example "bow harp". Hence, a much larger number of Wiktionary entries can be considered as instances of heteronymy, if one lexical item in a compound or in a phrasal entry is itself included in Wiktionary as a heteronym.

## 2 Targeted Lexical Databases

Although our current work is primarily intended at enriching WordNet, ultimately we aim at adding disambiguated pronunciation information to a series of lexical databases. Once the phonetic transcriptions are correctly stored in WordNet, this information can be propagated to BabelNet (Navigli and Ponzetto, 2012)[6] and all other lexical resources which are making use of WordNet.

### 2.1 Wordnets

As each WordNet is a sense inventory, it is particularly relevant to associate pronunciation information with the heteronyms it lists. Recently we witnessed the development of a new WordNet for English (McCrae et al., 2020), which is based on the Princeton WordNet (PWN, see (Fellbaum, 1998)), but aiming at an open source development policy. This makes this version of WordNet a good candidate for testing in a near future the addition of pronunciation information in a collaborative manner, using the corresponding GitHub platform.[7] The Open English WordNet (OEW) data can be downloaded in various formats, including XML, LMF[8] and RDF.

### 2.2 BabelNet

While BabelNet already combines wordnets and wiktionaries, as well as many other resources, it does not yet provide the phonetic transcription that it has extracted from various language versions of Wiktionary. Although BabelNet provides sound files in its word entries, those pronunciations are given by an external library that do not read from IPA codes. This library seems to be connected to the text-to-speech modules of the browser accessing the server, and utilises it to add pronunciation to some textual information on the BabelNet pages, like the entry and its associated definition(s) and example sentence(s).

Experimenting with BabelNet, we discovered that in fact a unique pronunciation for homographs is provided, leading thus to a number of wrong pronunciation examples. In this case we can see the importance of considering the IPA phonetic transcriptions for all senses of a heteronym. This way, the disambiguated IPA code of each sense could be used as input to the sound file generator of BabelNet. We hope that our work will prove beneficial in this endeavour.

### 2.3 ELEXIS – Dictionary Matrix

The Dictionary Matrix, under development within the ELEXIS project,[9] is a collection of linked dictionaries. The goal of this matrix is to enhance interoperability across resources and languages. For this, ELEXIS provides services for linking resources semi-automatically across languages at various matching levels such as headword, sense and lexeme. We plan to add pronunciation information to WordNet resources that are included in this linking exercise, as this can help in the particularly challenging sense linking task.

## 3 Our Approach

The first step of our work consisted in accessing the XML dump of the English Wiktionary resource,[10] and extracting from there, with the help of customised Python scripts, the pronunciation information associated with nouns, verbs, adjectives, and adverbs. As we can see in Figure 1, we also extracted the corresponding senses and associated examples sentences, as we need to keep the relation of

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

[7] Open English WordNet is accessible at https://github.com/globalwordnet/english-wordnet It is also accessible via a GUI: https://en-word.net/.

[8] LMF stands for "Lexical Markup Language", an ISO standard (Francopoulo et al., 2006), which has also be employed for encoding WordNet, as this is described for example in (Henrich and Hinrichs, 2010).

[9] https://elex.is/.

[10] The XML dumps of recent versions of the English edition of Wiktionary are available at https://dumps.wikimedia.org/enwiktionary/.

the pronunciation information with the corresponding meaning and the associated example sentences, if any is provided.

While we can report good progress in this task, there are still a few issues to solve, mainly due to the sometimes idiosyncratic way of encoding information in Wiktionary. While the overall XML structures of the lexical entries in Wiktionary is quite consistent, the linguistic information itself is encoded by making use of the Wiki mark-up language and with a number of options left to the (volunteering) encoders of the entries, so that extra lines of codes are necessary for dealing with those recurrent idiosyncratic cases. Still, we have extracted a large amount of lexical information that we have checked for validity. The numbers are given and discussed in the next section.

### 3.1 Some Figures

In this section we give some quantitative details on our current extraction work from Wiktionary.[11] A Wiktionary page is selected for processing if it contains within its English section one or more of the following Parts-of-Speech (PoS): noun, verb, adjective or adverb. This was the case for 829.342 Wiktionary pages, out of which the following lexical information was detected and extracted:

- nouns: 584.021
- verbs: 141.938
- adjectives: 139.887
- adverbs: 21.413
- pronunciation information for 72.067 entries (out of a total of 887.259 entries)

A note on the terminology is appropriate here. We call "Wiktionary pages" the Web resources that are accessed by a Wiktionary URL. So for the "lead" example, we access the Wiktionary page by typing "https://en.wiktionary.org/wiki/lead" in a browser. The element name "page" is in fact also used in the XML dump for marking an entry. A Wiktionary page typically covers more than one language (4 languages in our example). We are concentrating here on the English language, and in this case we see that 3 "etymologies" are listed, while two of them include the noun part-of-speech and all three include the verb part-of-speech. Those elements are the ones we call "entries" in the list of figures displayed just above.

On average, there is only 1,07 entries per English section in the selected pages. Many Wiktionary pages are about morphological variants of a lemma form, and those typically do not include PoS ambiguities. Therefore, we do not observe a significant amount of such PoS ambiguities in the English section of the total amount of selected Wiktionary pages, but there are many more ambiguities to be seen, if one concentrates on the Wiktionary pages that are leading to the lemma forms.

We observe that 815.192 English entries are without pronunciation information. Inspecting those, we see that in many cases the entries are in fact dealing with morphological variations (e.g. plural) of the ground form. In such cases we see the relatively straightforward possibility to automatically accommodate the pronunciation information of the lemma to the derived form. Also compound words are most often lacking the pronunciation information. An example of this is the adjective "leadlike". Although this would be more complicated, it could still be possible to derive the pronunciation of the compound word, as explained in the Future Work section.

We show the (shortened) output of our program for the extraction of nouns from the Wiktionary page "lead" in Figure 1.[12]

## 4 Formal Lexical Representation

In order to make the information we extracted from Wiktionary available in an interoperable and reusable format, we make use of the OntoLex-Lemon model, resulting from the W3C Community Group "Ontology Lexica" (Cimiano et al., 2016).[13] Figure 2 displays the general organisation of the core module of the OntoLex-Lemon model.

### 4.1 The RDF Encoding of the Open English WordNet

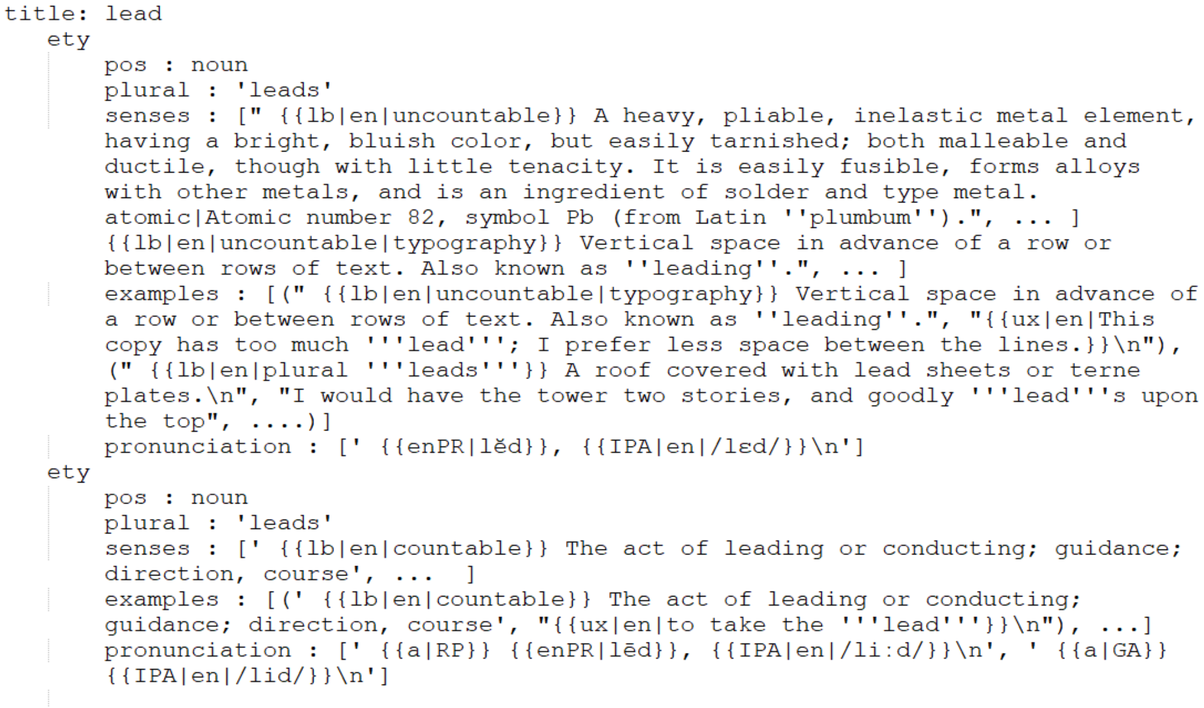
Our decision to use OntoLex-Lemon for representing the extracted lexical information from Wiktionary is also motivated by the fact that the Open English WordNet (OEW) has an export of its data in the so-called Global-Wordnet-RDF format,[14]

[11] We were using the XML dump of May 2020.

[12] At this stage of development, wiki mark-up signs are still included. In future versions, the data will be cleaned-up.

[13] See for more details https://www.w3.org/2016/05/ontolex/.

[14] For details and examples of the encoding, see http://globalwordnet.github.io/schemas/#rdf.



```
title: lead
   ety
       pos : noun
       plural : 'leads'
       senses : [" {{lb|en|uncountable}} A heavy, pliable, inelastic metal element,
       having a bright, bluish color, but easily tarnished; both malleable and
       ductile, though with little tenacity. It is easily fusible, forms alloys
       with other metals, and is an ingredient of solder and type metal.
       atomic|Atomic number 82, symbol Pb (from Latin ''plumbum'').", ... ]
       {{lb|en|uncountable|typography}} Vertical space in advance of a row or
       between rows of text. Also known as ''leading''.", ... ]
       examples : [(" {{lb|en|uncountable|typography}} Vertical space in advance of
       a row or between rows of text. Also known as ''leading''.", "{{ux|en|This
       copy has too much '''lead'''; I prefer less space between the lines.}}\n"),
       (" {{lb|en|plural '''leads'''}} A roof covered with lead sheets or terne
       plates.\n", "I would have the tower two stories, and goodly '''lead'''s upon
       the top", ....)]
       pronunciation : [' {{enPR|lĕd}}, {{IPA|en|/lɛd/}}\n']
   ety
       pos : noun
       plural : 'leads'
       senses : [' {{lb|en|countable}} The act of leading or conducting; guidance;
       direction, course', ...  ]
       examples : [(' {{lb|en|countable}} The act of leading or conducting;
       guidance; direction, course', "{{ux|en|to take the '''lead'''}}\n"), ...]
       pronunciation : [' {{a|RP}} {{enPR|lēd}}, {{IPA|en|/liːd/}}\n', ' {{a|GA}}
       {{IPA|en|/lid/}}\n']
```

Figure 1: The extracted information from the Wiktionary page "lead", focused on nouns, listing the PoS, the associated senses and examples, as well as the pronunciation belonging to each sense. (shortened)

which is using also the OntoLex-Lemon model. We display in the next 3 listings the way OEW is encoding information about "lead" in the Global-Wordnet-RDF format.[15] This representation is the one that will be used for automatically linking the disambiguated heteronym pronunciations to OEW.

In Listing 1 we see the way OEW encodes the original Princeton WordNet synset for the *metal* meaning of "lead".

```
pwnid:ewn-14667645-n
  owl:sameAs ili:i113959 ;
  wn:partOfSpeech wn:noun ;
  dc:subject ”noun.substance” ;
  wn:definition [ rdf:value
    ”a soft heavy toxic malleable
    metallic element; bluish white
    when freshly cut but tarnishes
    readily to dull grey”@en ] ;
  wn:hypernym pwnid:ewn-14649636-n ;
  wn:holo_substance pwnid:ewn-14700071-n ;
  wn:holo_substance pwnid:ewn-14694339-n ;
  wn:hyponym pwnid:ewn-14929227-n ;
  wn:hyponym pwnid:ewn-14929348-n ;
  wn:hyponym pwnid:ewn-15008253-n ;
  wn:hypernym pwnid:ewn-92464177-n ;
  a ontolex:LexicalConcept .
```

Listing 1: The Global-Wordnet-RDF representation of the Open English WordNet synset for the concept associated with *lead* in the *metal* sense (listing also semantic relations the synset is involved in)

Listing 2 below is displaying a meaning of "lead" that is a lexicalized sense of the synset introduced in Listing 1.

```
<#lead-ewn-14667645-n>
    ontolex:isLexicalizedSenseOf
        pwnid:ewn-14667645-n ;
    a ontolex:LexicalSense .
```

Listing 2: The Global-Wordnet-RDF representation of an OEW sense associated with the LexicalConcept pwnid:ewn-14667645-n

Listing 3 is then showing the OEW representation of the nominal lexical entry "lead", with all its senses.

```
<#lead-n>
  ontolex:canonicalForm [
    ontolex:writtenRep ”lead”@en
  ] ;
  ontolex:sense <#lead-ewn-05164526-n> ;
  ontolex:sense <#lead-ewn-14667645-n> ;
  ontolex:sense <#lead-ewn-05835238-n> ;
  ontolex:sense <#lead-ewn-01259362-n> ;
  ontolex:sense <#lead-ewn-13915822-n> ;
  ontolex:sense <#lead-ewn-06281532-n> ;
  ontolex:sense <#lead-ewn-13617665-n> ;
  ontolex:sense <#lead-ewn-10668135-n> ;
  ontolex:sense <#lead-ewn-08609721-n> ;
  ontolex:sense <#lead-ewn-06664322-n> ;
  ontolex:sense <#lead-ewn-06281845-n> ;
  ontolex:sense <#lead-ewn-05058239-n> ;
  ontolex:sense <#lead-ewn-03658258-n> ;
  ontolex:sense <#lead-ewn-03656591-n> ;
  ontolex:sense <#lead-ewn-03656410-n> ;
  ontolex:sense <#lead-ewn-03610056-n> ;
```

[15] The encodings are taken from https://en-word.net/lemma/lead.

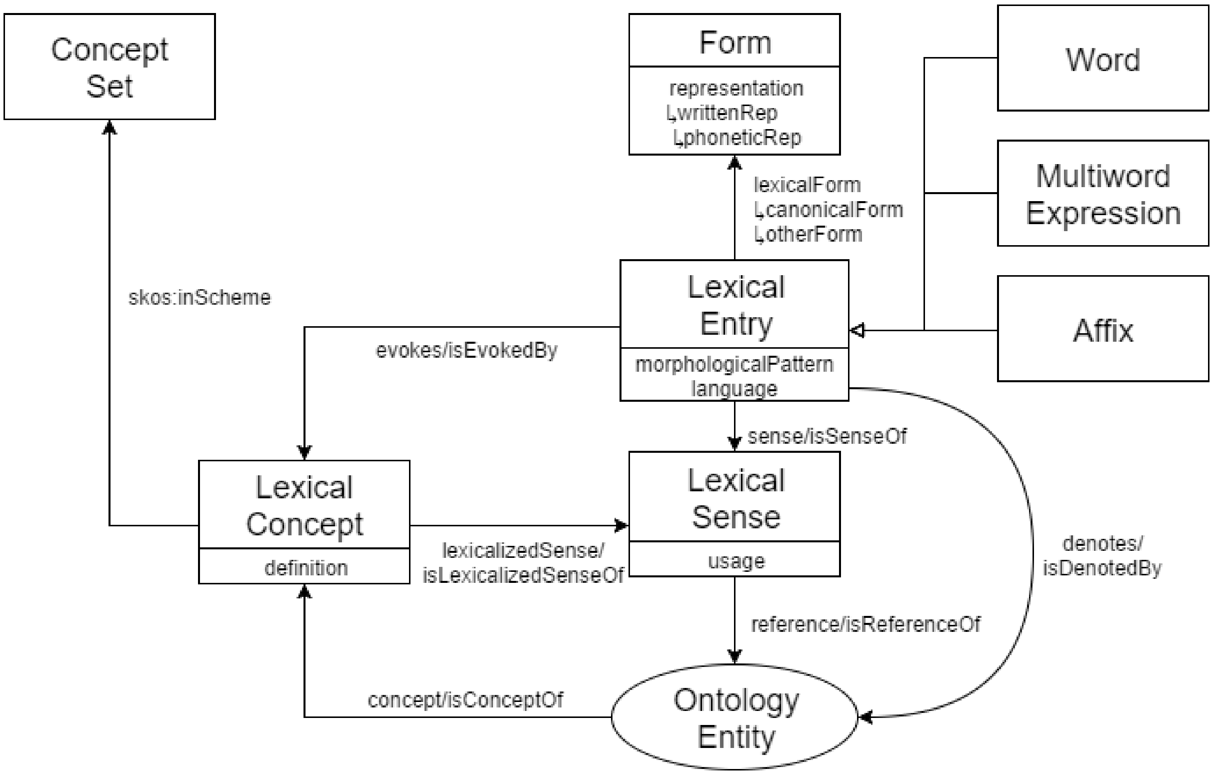


Figure 2: The core module of OntoLex-Lemon. Graphic taken from https://www.w3.org/2016/05/ontolex/.

```
  ontolex:sense <#lead−ewn−01258857−n> ;
  wn:partOfSpeech wn:noun ;
  a ontolex:LexicalEntry .
```

Listing 3: The Global-Wordnet-RDF representation of the OEW entry “lead”

In this representation, the canonical form is included as the value of a blank node that just gives information about its written representation. We aim at adding the phonetic representation. But as not all the senses listed in this entry are related to the same concept, we can not assume one canonical form with the same pronunciation for all senses, and we have to depart from the modelling displayed in Listing 3.

### 4.2 Adapting the Representation

In this section we present the current OntoLex-Lemon representation we suggest for elements of the lexical information extracted from Wiktionary, for the example of “lead”, in its *metal* meaning.

Listing 4 is just displaying the Lexical Concept representation for “lead”, similar in part to the representation shown in Listing 1, but without semantic relations. A major difference is that the definition is now “outsourced”, as we introduce definitions as instances of a class “:Definition”, as can be seen in Listing 5. We are also adding a link to a Wikidata page.

```
:LexicalConcept_1
  rdf:type ontolex:LexicalConcept ;
  rdfs:label ”\”lead\””@en ;
  skos:definition
    :Definition_Concept_1_English_Lead_1 ;
  skos:topConceptOf :ConceptSet_1 ;
  ontolex:isConceptOf
    <https://www.wikidata.org/wiki/Q708> ;
  ontolex:isEvokedBy :lex_lead_1 ;
  ontolex:lexicalizedSense :sense_lead_1 ;
.
```

Listing 4: Our suggested OntoLex-Lemon representation for the OEW entry “lead”

```
:Definition_Concept_1_English_Lead_1
  rdf:type :Definition ;
  rdfs:label ”\”A heavy, pliable,
  inelastic metal element, having
  a bright, bluish color, but easily
  tarnished; both malleable and
  ductile, though with little tenacity.
  It is easily fusible, forms alloys
  with other metals, and is an ingredient
  of solder and type metal. Atomic number
  82, symbol Pb (from Latin plumbum). ;
.
```

Listing 5: A Wiktionary definition for “lead” as an instance of the class “:Definiton”

Listing 6 introduces one sense for the *metal* meaning of “lead” in Wiktionary.

```
:sense_lead_1
  rdf:type ontolex:LexicalSense ;
  rdfs:label ”\”lead\””@en ;
  ontolex:isLexicalizedSenseOf
    :LexicalConcept_1 ;
  ontolex:isSenseOf :lex_lead_1 ;
  ontolex:reference
    <https://www.wikidata.org/wiki/Q708> ;
  ## ontolex:usage lexinfo:singular ;
.
```

Listing 6: Introducing a sense for on the meanings of “lead” in Wiktionary

The reader can see that we link this sense to a specific lexical entry for “lead”, as we have now two entries for this word. The commented line “##ontolex:usage lexinfo:singular” shows the possibility to express that this sense requires the word to be used in singular. But we disregard this encoding here, as we are introducing also different forms for the noun “lead”, one per pronunciation. One case is shown in Listing 7.

```
:lex_lead_1
  rdf:type ontolex:Word ;
  lexinfo:partOfSpeech
     lexinfo:noun ;
  rdfs:label ”\”lead\””@en ;
  ontolex:canonicalForm
     :form_lead_singular_1 ;
  ontolex:evokes :LexicalConcept_1 ;
  ontolex:otherForm
     :form_lead_plural_1 ;
  ontolex:sense :sense_lead_1 ;
.
:form_lead_singular_1
  rdf:type ontolex:Form ;
  lexinfo:number lexinfo:singular ;
  rdfs:label ”\”lead\””@en ;
  ontolex:phoneticRep
    ”\textipa{[/lEd/]}/en–GB–fonipa” ;
  ontolex:writtenRep ”\”lead\””@en ;
.
```

Listing 7: The specific lexical entry and its related form – with the pronunciation information

Related conclusive experiments were also done for encoding lexical information extracted from the German Wiktionary (Declerck et al., 2020). It is suggested in (Declerck et al., 2020) that one could link specific senses of an entry to a lexical form carrying a specific pronunciation (by the use of the ontolex:phoneticRep) by applying restrictions that are defined in the lexicog module ((Bosque-Gil et al., 2019)[16] of OntoLex-Lemon. However, in our current experiment, we think that it might be more effective to just duplicate the lexical forms along the line of their pronunciation (even if they have the same gender and number features), and to point to those from the lexical sense via the corresponding lexical entry.

## 5 Sense Linking

In the following phase of our work, we plan to connect the extracted information with the correct WordNet synsets. After extracting the pronunciation information from Wiktionary, the subsequent step of our work lies in sense disambiguation and linking. More specifically, this task requires correctly inferring which of the heteronym synsets is the right match for the pronunciation information we have extracted from the Wiktionary entry. In order to disambiguate the word sense, we can utilize the WordNet synsets of the heteronymous senses as well as their description and examples from Wiktionary.

Our initial approach relies on comparing the document similarity between WordNet synsets and the matching Wiktionary entries. Firstly, we create ’documents’ by concatenating the definitions and examples for all the senses of the ambiguous word. In the case of “lead”, we have decided to combine all the possible sub-senses using their PoS tag. In this way, according to WordNet, we end up with two broader senses for “lead”: a broad noun sense and a broad verb sense. These two documents need to be compared with the two documents extracted from Wiktionary, using the same approach. After tokenization, punctuation cleaning and stopwords removal, document similarity is calculated using TFIDF and the bag-of-words approach. For this purpose we have utilized the Docsim library from Gensim[17].

The preliminary work shows promising results. In Table 1 we can see these similarity comparisons for the example word “lead”. Columns represent the sense documents extracted from Wiktionary, represented by their pronunciations, while rows represent the senses extracted from WordNet. The highest similarity scores are for the correct combinations of senses, which is the outcome we would expect. We believe that this approach, with modifications, can be used for automatic heteronym sense linking on a greater scale. However, joining all the sub-senses is certainly not the best solution.

[16] See https://www.w3.org/2019/09/lexicog/ for more details of the specifications of the module.

[17] The Docsim library is explained here: https://radimrehurek.com/gensim/similarities/docsim.html

| IPA code | [lɛd] | [liːd] |
|---|---|---|
| lead.noun | 0.4272 | 0.0672 |
| lead.verb | 0.2176 | 0.4581 |

Table 1: Similarity scores for sense matching

The noun sense of the word lead can also refer to an advantage held by a competitor in a race, in which case the correct pronunciation is the second one. So we can see that sense granularity is also an important aspect when it comes to heteronym disambiguation.

## 6 Related Work

Our work with the English Wiktionary is an extension and a refinement of a first experiment dealing with the German version of Wiktionary, with the aim of enriching a new open WordNet for German with pronunciation information (Declerck et al., 2020). In both cases, we make use of the OntoLex-Lemon community standard for encoding the heteronyms (and other entries). Our current development is aiming at including the results into various integrated or interlinked lexical databases. We are also aiming at automatically adding pronunciation information to derived terms, on the base of sense-linking algorithms.

The work presented in (Declerck, 2020) describes an approach for linking the Open Dutch WordNet to external lexical resources, including the Dutch version of Wiktionary, with the goal of enriching the lemmas in the WordNet entries with morphological variants. But the work was not dealing with pronunciation information.

(Schlippe et al., 2010) assess the quality of pronunciation information in Wiktionary for four languages (English, French, German, and Spanish) and come to satisfying results, especially in the case of French, when it comes to the evaluation of the coverage and also to the impact on automatic speech recognition (ASR) systems, especially in the case of Spanish. This already older study comforted us in the opinion that extracting pronunciation information from Wiktionary can deliver a relevant source of data for our experiment consisting in equipping wordnets with pronunciation information.

In recent years, relevant research regarding heteronyms is done in the field of speech synthesis. For example, the work of (Samsudin and Rahim, 2019) focuses on handling heteronym ambiguity for a text-to-speech (TTS) system for Malay language. Although there are only 12 unique heteronyms in Malay, this research emphasises the importance of conducting a specific study on heteronym words and their pronunciation by TTS systems. Other important work in this field includes the patents of (Henton and Naik, 2014) and (Wang et al., 2011). Both models focus on heteronym pronunciation for dialogue systems, using the user's input to correctly predict the pronunciation of the output heteronym.

## 7 Future Work

A crucial phase of the future work involves evaluation. For this we could use some existing dictionaries which contain pronunciation information. Since pronunciation is an inevitable part of translation dictionaries, extracting the information from such sources could substantially enlarge the underlying resource and also serve as a basis for evaluation.

One interesting possibility for an expansion of the scope of this work can be found in compound words and derived terms. After correctly disambiguating the heteronymous lemma, we can use this information to produce the IPA of the compound words which contain it. This would be done following the rules of metrical phonology (Kreidler, 2004). If that would prove too complex we could produce the IPA without stress information. We could also use etymology information from Wiktionary to produce pronunciation descriptions for compounds.

## Acknowledgements

Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number 825182. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015. The work described in this paper was also pursued in part in the larger context of the COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science". We also thank the anonymous reviewers for their insightful comments.

## References

Julia Bosque-Gil, Dorielle Lonke, Jorge Gracia, and Ilan Kernerman. 2019. Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.*, pages 726–746, Brno, Czech Republic. Lexical Computing CZ s.r.o.,.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report.

Thierry Declerck. 2020. Towards an extension of the linking of the open dutch wordnet with dutch lexicographic resources. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 33–35. ELRA.

Thierry Declerck, Lenka Bajcetic, and Melanie Siegel. 2020. Adding pronunciation information to wordnets. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. ELRA.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Verena Henrich and Erhard W. Hinrichs. 2010. Standardizing wordnets in the ISO standard LMF: wordnet-lmf for germanet. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 456–464. Tsinghua University Press.

Caroline Henton and Devang Naik. 2014. Disambiguating heteronyms in speech synthesis.

Charles Kreidler. 2004. *Prefixes, Compound Words, and Phrases*, chapter 12. John Wiley Sons, Ltd.

M. Martin, G.V. Jones, and D.L. Nelson. 1981. Heteronyms and polyphones: Categories of words with multiple phonemic representations. *Behavior Research Methods & Instrumentation*, 13:299–307.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019a. English wordnet 2019 – an open-source wordnet for english. In *Proceedigns of the 10th Global Wordnet Conference*. Global Wordnet Association. To appear.

John P. McCrae, Carole Tiberius, Anas Fahad Khan, Ilan Kernerman, Thierry Declerck, Simon Krek, Monica Monachini, and Sina Ahmadi. 2019b. The elexis interface for interoperable lexical resources. In *Proceedings of the eLex 2019 conference*, pages 642–659. CELGA-ILTEC, University of Coimbra, Lexical Computing CZ, s.r.o.

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, MMW@LREC 2020, Marseille, France, May 2020*, pages 14–19. The European Language Resources Association (ELRA).

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.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

N. Samsudin and L. N. Rahim. 2019. Rapid heteronym disambiguation for text-to-speech system. In *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6.

Tim Schlippe, Sebastian Ochs, and Tanja Schultz. 2010. Wiktionary as a source for automatic pronunciation extraction. In *11th Annual Conference of the International Speech Communication Association, Makuhari, Japan*. Interspeech 2010.

Xi Wang, Xiaoyan Lou, and Jian Li. 2011. Speech synthesis with fuzzy heteronym prediction using decision trees.