ENGLAWI: From Human- to Machine-Readable Wiktionary

Franck Sajous, Basilio Calderone, Nabil Hathout

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3016–3026
Marseille, 11–16 May 2020

ENGLAWI: From Human- to Machine-Readable Wiktionary

Franck Sajous, Basilio Calderone and Nabil Hathout
CLLE-ERSS, CNRS & Université de Toulouse 2
{franck.sajous, basilio.calderone, nabil.hathout}@univ-tlse2.fr

Abstract

This paper introduces ENGLAWI, a large, versatile, XML-encoded machine-readable dictionary extracted from Wiktionary. ENGLAWI contains 752,769 articles encoding the full body of information included in Wiktionary: simple words, compounds and multiword expressions, lemmas and inflectional paradigms, etymologies, phonemic transcriptions in IPA, definition glosses and usage examples, translations, semantic and morphological relations, spelling variants, etc. It is fully documented, released under a free license and supplied with G-PeTo, a series of scripts allowing easy information extraction from ENGLAWI. Additional resources extracted from ENGLAWI, such as an inflectional lexicon, a lexicon of diatopic variants and the inclusion dates of headwords in Wiktionary's nomenclature are also provided. The paper describes the content of the resource and illustrates how it can be – and has been – used in previous studies. We finally introduce an ongoing work that computes lexicographic word embeddings from ENGLAWI's definitions.

Keywords: Wiktionary, Machine-Readable Dictionary of Contemporary English, Definitions-based Embeddings

1. Introduction

This paper introduces ENGLAWI, a structured and normalized version of the English Wiktionary encoded into a workable XML format. ENGLAWI is freely available for download,[1] fully documented, and is supplied with G-PeTo (GLAWI Perl Tools), a set of scripts that helps extract specific information from GLAWI dictionaries. Contrary to other approaches designed to extract particular data from Wiktionary, ENGLAWI is part of a series of works aiming to provide the full body of information encoded in the collaborative dictionaries. In previous papers, we presented GLAWI (Sajous and Hathout, 2015; Hathout and Sajous, 2016), extracted from the French Wiktionnaire and GLAWIT (Calderone et al., 2016), extracted from the Italian Wikizionario. In these papers, we described the conversion and the standardization of the heterogeneous data extracted from the dictionaries' dumps and we illustrated how specific lexicons can be easily tailored on demand. We did adopt an identical approach to develop ENGLAWI. The current paper describes the content of ENGLAWI, illustrates how it is structured and suggests possible uses for linguistic studies and NLP applications.

The remainder of the article is organized as follows. In Section 2, we discuss the usefulness of machine-readable dictionaries (MRD) for linguistics and NLP. We compare in Section 3 our approach to related works on information extraction from Wiktionary. Then, we describe the structure of ENGLAWI in Section 4 and we introduce additional tools and resources provided with ENGLAWI in Section 5. Finally, we recap in Section 6 how ENGLAWI, GLAWI and GLAWIT have been used in various linguistic and NLP works and we propose a new method for computing lexicographic word embeddings from ENGLAWI's definitions.

2. Are MRD of any use for NLP?

In a six-week online course on English dictionaries running in September and October 2019 (Creese et al., 2018), lexicographers explored "the place of dictionaries in the modern world" and raised the question: "Would it matter if there were no longer any dictionaries?". The question stands for humans. A parallel question could be: are MRD relevant – and to which extent – for NLP? Back in the nineties, Ide and Veronis (1993) analyzed the reasons of disappointing results of extracting lexical and semantic knowledge from MRD in a paper entitled "Extracting knowledge bases from MRD: Have we wasted our time?". Ide and Veronis formulated two postulates: 1) MRD contain information that is useful for NLP and 2) this information is relatively easy to extract from MRD. After reviewing various studies, they observed that (in 1993) "the 15 years of work has produced little more than a handful of limited and imperfect taxonomies". One explanation they gave is that dictionary information was flawed. The authors also wondered if extracting information from MRD was as simple as applying strategies described in the papers they reviewed. The subsequent rise of corpus linguistics and the ever-growing use of machine-learning over large corpora could make MRD a thing of the past. Two counter-arguments can be made. First, works surveyed by Ide and Veronis were based on dictionaries from publishing houses that were old enough to have fallen into the public domain. The availability of a free, huge and ongoing (updated) dictionary such as Wiktionary could make a difference. Second, lexical knowledge may be still necessary for some tasks or may at least improve systems trained on corpora. Two decades after the study by Ide and Veronis, Gurevych et al. (2016) write in the foreword of a book on linked lexical knowledge bases that lexical semantic knowledge is vital for most tasks in NLP. Later on, they also write that the benefit of using lexical knowledge bases in NLP systems is often not clearly visible (Gurevych et al., 2016, p. 65–66), leaving the reader in two minds. In several works however, information extraction from Wiktionary has proven successful. For example, Schlippe et al. (2010) created pronunciation dictionaries from Wiktionary within a speech recognition and synthesis process. The authors assessed the good coverage and quality of the extracted transcriptions. De Smedt et al.

[1] Resources and tools presented in the paper are available from `redac.univ-tlse2.fr/lexicons/englawi/`

(2014), noting that basic resources (especially free ones) were still missing for many languages, including Italian, developed a weakly-supervised, fast, free and reasonably accurate tagger for Italian, created by mining words and their part-of-speech from Wiktionary. Ács (2014) applied a triangulation method to Wiktionary to create a pivot-based multilingual dictionary. Metheniti and Neumann (2018) produced inflectional paradigms for over 150 languages from Wiktionary. Segonne et al. (2019) investigated which strategy to adopt to achieve WSD for languages lacking annotated sense disambiguated corpora (i.e. languages other than English). Focusing on verb disambiguation in French, they resorted to the sense inventory and to the manually sense tagged examples extracted from the French Wiktionnaire to train WSD systems. All these works, like other successful use of data extracted from Wiktionary, presented in Section 6, confirm that Wiktionary-based MRD may benefit NLP tools and linguistic studies.

3. Related works on information extraction from Wiktionary

Ide and Veronis (1993) questioned the simplicity of extracting information from MRD (cf. Section 2). For different reasons, extracting information from Wiktionary is not as easy as often described. We explained in (Sajous and Hathout, 2015) and (Hathout and Sajous, 2016) that the wikicode (Wiktionary's underlying syntax) heavily relies on the use of templates that need to be analyzed and reimplemented to fully and faithfully extract Wiktionary's content. We also noted in these papers that the syntax highly differs from a language edition to another. Several authors focus on the extraction of specific information from Wiktionary. Because partial and noisy data meet their needs, they do not report such difficulties. Most of them use the English dump to extract information in another target language and purely ignore the edition of Wiktionary in the language they are working on. We analyze below the pros and cons of different approaches to extracting data from Wiktionary.

First pioneering work on using Wiktionary for NLP was led by Zesch et al. (2008) for semantic relatedness computation. The authors released JWKTL, an API giving access to the data of the English and German wiktionaries. Navarro et al. (2009) worked on synonymy mining and made available the first versions of the English and the French wiktionaries as MRD, called WiktionaryX. An advantage of JWKTL is that a user can download a recent dump of Wiktionary and access its current data. A drawback is that it handles regular wiki markups such as hyperlinks, bold and italic, but do not adequately process all the MediaWiki templates. For example, the linguistic labels found in definitions are ignored[2] and nested templates are not correctly processed.[3] As a result, asking for the plain text of a sense of the adjective *sweet*, whose wikicode is:

```
{{lb|en|informal|followed by {{m|en|on}}}}
[[romantic|Romantically]] [[fixate|fixated]],
[[enamor|enamoured with]], [[fond|fond of]]
```

produces:

```
}} Romantically fixated, enamoured with, fond of
```

instead of:

```
(informal, followed by on) Romantically fixated,
enamoured with, fond of
```

Resources such as WiktionaryX have the advantage of being ready to use, but their content age if they are not updated. Sérasset (2012), whose aim was to build a multilingual network based on Wiktionary, wrote that he focused on *easily extractable* entries. The then resulting graph for French, including 260,467 nodes, was far from exhaustive. The author, however, did not purport its resource to be exhaustive: in (Sérasset, 2015), he wrote that the main goal of his efforts was not to extensively reflect the content of Wiktionary. But the main drawback of extractors that overlook the wikicode complexity, and especially the importance of handling templates correctly, is not the lower amount of data retrieved. It is rather that they generate ill-extracted data. For example, 9% of DBnary's definitions are empty, and some others are not consistent with Wiktionary's content (cf. Section 4.6.1.). Hellmann et al. (2013) created a declarative syntax attempting to enable non-programmers to tailor and maintain extraction tools, and to easily adapt existing extractors to another Wiktionary language edition. The approach they proposed may be sufficient to extract translations or semantic relations and output RDF triples. However, the authors overlooked the differences between Wiktionary's language editions and the importance of templates reimplementing for some languages. To our knowledge, no such method allows to produce, for example, clean definitions from wiktionaries such as the French one (Wiktionnaire). Babelnet (Navigli and Ponzetto, 2010), which mixes lexicographic and encyclopedic knowledge, gives, for English words, definitions taken from Wikipedia, WordNet and Wiktionary. For French words, definitions are mostly taken from Wikipedia and never from Wiktionnaire. The reason may be that the information from the French dictionary is harder to parse. Thus, *alpabète* 'literate' is only defined by an equivalent, *lettré*, while it has a full definition in Wiktionnaire. The verb *divulgâcher* 'spoil (reveal the ending)', missing from Wikipedia, is also missing from Babelnet, though it has a definition in Wiktionnaire. Another side effect is that, the titles of Wikipedia's articles being mostly nouns, words of other parts of speech are not well covered by Babelnet when they are also absent from WordNet. The definition of *consensuel* 'consensual' is taken from WordNet and given in English (*existing by consent*) where it could be defined in French by *issu d'un consensus*, taken from Wiktionnaire. The equivalent Italian adjective *consensuale* is defined by the same wording as the noun *consenso* 'consensus', although a definition exists in the Italian wiktionary (*che si fa col consenso della o delle altre parti*).

The most comparable works to ours are knoWitiary (Nastase and Strapparava, 2015) and IWNLP (Liebeck and Conrad, 2015). Nastase and Strapparava's purpose was "to obtain a coherent and consistent lexical resource that contains as much information as possible about words and their relations". Liebeck and Conrad developed a lemmatizer for German and focused on the extraction of inflections for this language. They insisted on the importance of templates reimplementation for that purpose. Both papers

[2] The `getMarker()` method always returns an empty string (with JWKTL 1.1.0 and Wiktionary's 2019-11-01 dump).

[3] Corresponding text is removed and extra brackets remain.

compared the extracted data to that obtained with JWKTL and revealed lacks in the latter. Unlike IWNLP, whose both source code and data are made available for download, knoWitiary is not publicly released. Another original method is that of Kirov et al. (2016), who relied on the analysis of HTML pages of different language editions of Wiktionary to extract morphological paradigms in various languages, with minimal language-specific tuning. According to the authors, their approach achieved – for 3 tested languages – results comparable in quality and quantity to that obtained with previous, fine-tuned methods. Extracting data in 350 languages, the authors concluded their method contributes a uniquely large morphological resource, they called UniMorph. ENGLAWI did not exist by that time. Resources existed however for French, namely GLÀFF (Sajous et al., 2013a; Hathout et al., 2014b) and GLAWI (Sajous and Hathout, 2015; Hathout and Sajous, 2016), that achieved a better coverage (cf. Section 4.2.), but were not taken into account in the comparison led by Kirov et al. (2016).

What we propose in the current paper is a resource for a single language – English – containing the full body of information encoded in Wiktionary, including notably definitions and etymologies, as clean as possible and as conform as possible to Wiktionary's original content. ENGLAWI, we hope, will help conduct further research without the need of wikicode parsing.

4. Resource description

The general structure of an article is depicted in the Figure 1, that illustrates the encoding of the article *frog*.

4.1. An ad hoc format

When working within the TEI Dictionary Working Group, Ide and Véronis (1995) noted that dictionaries were among the most complex text types treated in the TEI, that the structure of dictionary entries was highly variable, both within and among dictionaries and that any piece of information can go anywhere in some dictionary: *"In large, complex dictionaries such as the OED, unusual exceptions [...] are fairly common. As a result, it is probably impossible to define a fixed structure"*. In order to handle this situation, a new element (`entryFree`) was added to the DTD under development, allowing any component of the dictionary to be combined in any order or organization. Two decades later, Bański et al. (2017) review the results achieved in the context of TEI-Lex0. They note that the TEI guidelines, aiming at being able to encode any existing work, provide multiple encoding solutions and have been criticized for being too complex. To *"secure interoperability"*, a strategy has been to provide a format *"that may not be able to handle all the potential variation, but will instead address 90% of the phenomena, 90% of the time"*. In short: to ensure interoperability, reduce the amount of data to be made interoperable. In parallel, Romary et al. (2019) initiated an *"in-depth review"* of LMF, the "*de jure standard which constitutes a framework for modeling and encoding lexical information*". According to the authors, the goal is to create a more modular, flexible

```
<article>
  <pageId>39323</pageId>
  <title>frog</title>
  <meta>
    <category>English ethnic slurs</category>
    <category>English informal demonyms</category>
    <category>Amphibians</category>
    <reference>Webster 1913</reference>
  </meta>
  <text>
    <pronunciations>
      <pron area="UK">fɹɒg</pron>
      <pron area="US">fɹɑg</pron>
    </pronunciations>
    <etymologies>
      <etymology nb="1">From Middle English frogge, from
      Old English frogga, frocga (frog)...</etymology>
      <etymology nb="2">From frog legs, stereotypical
      food of the French. Compare rosbif...</etymology>
      <!-- ... -->
    </etymologies>
    <pos type="noun" lemma="1" etymNb="1">
      <paradigm>
        <inflection gracePOS="Nc-s" form="frog"/>
        <inflection gracePOS="Nc-p" form="frogs"/>
      </paradigm>
      <definitions>
        <definition level="1"><txt>A small tailless
        amphibian of the order Anura...</txt></definition>
        <!-- ... -->
      <definitions>
      <section type="morpho">
        <item type="derived">froggery</item>
        <item type="derived">froggish</item>
        <!-- ... -->
      </section>
      </translations>
        <trans lang="af">padda</trans>
        <trans lang="fi">sammakko</trans>
        <!-- ... -->
      </translations>
    </pos>
    <pos type="verb" reg="0" lemma="1" etymNb="1">
      <!-- ... -->
    </pos>
    <pos type="noun" lemma="1" etymNb="2">
      <definitions><definition level="1"><labels>
          <label type="attitudinal" value="offensive"/>
        </labels>
        <txt>A French person.</txt></definition>
        <definition level="1"><labels>
          <label type="diatopic" value="Canada"/>
          <label type="attitudinal" value="offensive"/>
        </labels> <txt>A French-speaking person
        from Quebec.</txt> </definition>
      </definitions>
      <section type="semRel">
        <item type="synonym">French person</item>
      </section>
    </pos>
    <!-- ... -->
  </text>
</article>
```

Figure 1: General structure of an article: *frog* (excerpt)

and durable follow up to the original LMF standard published by ISO in 2008, judged too rich and too complex. LMF provides standard solutions to encode bricks of lexical knowledge rather than dictionaries taken as a whole. Nevertheless, one can wonder how, since LMF is a metamodel, each of which could give a particular instantiation, this standard, even "reloaded", can guarantee interoperability. For languages that lack resources (e.g. MRD), such as French or Italian, interoperability is not an issue, however: let us recall the truism that the question of interoperability arises when several resources exist. Another issue raised by Nastase and Strapparava (2015) when discussing resources mapping (namely Wiktionary and WordNet) also applies to the process of making resources fit into norms: when a re-

source (e.g. Wiktionary) provides different types of information compared to others (e.g. WordNet), it implies that much is discarded when doing a mapping. Similarly, an attempt to encode the unique knowledge that idiosyncratic resources such as Wiktionary provide will result in discarding "unorthodox" lexical information. Unfortunately, in Wiktionary, the exception is the rule. Conversely, when developing DBnary, Sérasset (2015) decided to encode several language editions of Wiktionary into the LEMON model, although he judged this model not sufficient to represent lexical data that are currently available in dictionaries. Moreover, LEMON, he wrote, assumes that all data is well-formed and fully specified, which is not the case in Wiktionary – and, according to Ide and Véronis (1995), neither is the case case of "regular" dictionaries. As a consequence, Sérasset had to extend the LEMON model to encode Wiktionary's data. And even then, the extended model does not take into account, for example, Wiktionary's nested definitions (cf. Section 4.6.1.). Thus, while standard formats exist for dictionary encoding, we designed an ad hoc structure to encode ENGLAWI. As explained above, we wanted ENGLAWI's content to be as close as possible to the one of Wiktionary. Instead of twisting Wiktionary's content to make it fit into a given standard – or twisting any standard so as to accommodate Wiktionary's content –, we decided to model ENGLAWI's macro- and micro-structure so that it sticks to that of Wiktionary. Besides conforming to Wiktionary's content, ENGLAWI's structure is also quasi-identical to that of GLAWI and GLAWIT. The similar structure of the English, French and Italian resources makes it *really* easy to adapt a tool designed for a resource to another one (unlike the adaptation of a wiki extractor from a language to another, whose difficulty is often underestimated, as we said in Section 3). Moreover, as its format makes information extraction very simple, any user interested in producing, for instance, data in a given RDF standard, can easily write an extractor for that purpose.

4.2. Nomenclature

Wiktionary's basic unit is the written form, associated with a given page (bound to a given URL). Accordingly, ENGLAWI's articles correspond to a given written form. Both lemma and inflected forms may appear as headwords in Wiktionary. When several parts of speech (POS) or homographs correspond to the same written form, the corresponding article contains one separate POS section for each one of them. Each POS section includes definitions (glosses and examples), and several optional subsections described in Section 4.6. The Table 1 gives the number of entries per POS for lexical words, divided into lemmas (e.g. *deal*) and non lemmas (e.g. *dealt*). The columns entitled *lemmas* and *non lemmas* correspond to Wiktionary's headwords, i.e. words having a dedicated page in Wiktionary. Inflections can appear as headwords. They may also be found in articles' microstructure. The number of inflected forms included in inflectional paradigms (see Section 4.6.3. for details on their construction), as well as the number of senses per POS (lemma) is also given.

POS	Headwords		Senses	Paradigms' inflections
	Lemmas	Non lemmas		
Common N.	282,912	196,790	361,865	511,857
Verb	36,930	80,965	62,797	493,378
Adjective	108,760	4,418	130,025	114,554
Proper N.	46,260	9,936	60,921	91,122
Adverb	18,059	155	20,287	18,616

Table 1: ENGLAWI's nomenclature

POS	Lemmas			Inflections		
	UM1	DBnary	ENGLAWI	UM1	DBnary	ENGLAWI
Noun	159,917	218,218	282,912	166,314	228,407	511,857
Verb	23,532	36,222	36,930	73,185	114,183	493,378
Adj.	52,552	52,744	108,760	85,955	106,223	114,554
Adv.	-	10,992	18,059	-	21,968	18,616

Table 2: ENGLAWI's nomenclature compared to that of UniMorph 1.0 (UM1) and DBnary

The Table 2 compares ENGLAWI's nomenclature (based on a 10/2017 dump) to that of UniMorph 1.0 (based on a 06/2015 dump), and DBnary (2020-02-22 release). The figures given for UniMorph are taken from the (Kirov et al., 2016) paper. The UniMorph 2.0 data, available for download (`http://unimorph.org/`), produces slightly different results (e.g. 22,766 verbs for English, other POS not being available for this language). The difference of size between ENGLAWI and UniMorph cannot be explained only by a growth of Wiktionary in a two-year period. We think that approaches like (Kirov et al., 2016) are relevant as far as a highly multilingual resource is desired. When working on a single target language, "finely-tuned" extractions as ours provide better results (another difference is that we use the French dump to produce data for French and the English one to generate data for English). This is confirmed when comparing the French version of UniMorph to GLAWI's nomenclature: both resources have been produced with dumps released within a 6-month timespan (June-December 2015) but produce totally different results: GLAWI includes 7.3x more nouns than the French version of UniMorph, 3.3x more verbs and 5.8x more adjectives. We also obtain better results regarding the size of inflectional paradigms. DBnary's number of lemmas are taken from the resource's *core data*, available as a turtle file. The concept of "lemma" is however somewhat dubious in this file: inflected forms are considered canonical forms of lexical entries (e.g. both *children* and *child*, or *cats* and *cat* are "canonical forms", not related to each other in the core resource). ENGLAWI's numbers of inflections are also superior to that of knoWitiary, announced in (Nastase and Strapparava, 2015). However, results are subject to discussion because they depend, for example, on how many forms are expected to be found in verbal paradigms. Too few is said in the corresponding paper to conduct a proper comparison and, unlike UniMorph, knoWitiary is not publicly available. Regarding the number of verbal inflections, we probably extract more data, but the difference also stems from our extensive description of paradigms (cf. Section 4.6.3.).

4.3. Metadata

The `metadata` section contains categories and references. Just like in Wikipedia, categories are manually assigned to

pages in Wiktionary. Category attribution may also result from the use of linguistic labels in the articles. References are used by contributors to cite or refer to external sources. Such sources may be online or printed dictionaries, specialized websites, etc. The Figure 1 shows that the article *frog* belongs to three categories (*English ethnic slurs*, *English informal demonyms* and *Amphibians*) and refers to the 1913 edition of the *Webster's Revised Unabridged Dictionary of the English Language*.

4.4. Etymologies

In case of multiple etymologies, each `etymology` tag is numbered and POS sections refer to a given etymology (cf. Section 4.6.), as in the article *frog* (Fig. 1). Like in GLAWI, etymologies and definitions are available in four formats: wikicode, XML, plain text and syntactic parsing (more details in Section 4.6.1.). An illustration of an etymology of the word *cat* is given in Figure 2. Diachronic information may be provided: etymons and languages, as well as relations between etymons (cognate, borrowing, derivation, inheritance and calque).
Morphological information on word formations (affixes, compounds, derivations, etc.) also occurs in etymologies, as illustrated in Figure 3. This data is usually more reliable than underspecified relations (*derived* and *related*, that contributors use inconsistently) found in Wiktionary's sections entitled *derived terms* (cf. Section 4.6.2.). Such information provides material to be included in morphological resources such as Démonette (Hathout and Namer, 2014).

```
<etymology nb="1">
  <wiki>From {{inh|en|enm|cat}}, {{m|enm|catte}},
       from {{inh|en|ang|catt||male cat}}...
  <xml>From <etym type="inherited"><lang langCode="enm">
    Middle English</lang> <etymon>cat</etymon></etym>,
    <etymon langCode="enm" langName="Middle English">
    catte</etymon>, from <etym type="inherited">
    <lang langCode="ang">Old English</lang>
    <etymon gloss="male cat">catt</etymon></etym>...</xml>
  <txt>From Middle English cat, catte,
       from Old English catt ...</txt>
  <parsed> <!-- ... --> </parsed>
</etymology>
```

Figure 2: One shortened etymology (out of six) for *cat*

```
<xml>From
     <wordFormation type="affix">multicultural|-ism
     </wordFormation></xml>
<xml><wordFormation type="compound">
    news|letter</wordFormation></xml>
<xml><formOf type="clipping">potentiometer
    </formOf>.</xml>
```

Figure 3: Word formations in *multiculturalism*, *newsletter* and *pot* etymologies

4.5. Pronunciations

Pronunciation sections provide IPA transcriptions of words. They may take into account diachronic variation, as illustrated in Figure 1, where the different transcriptions of *frog* are given for both the UK and the US pronunciations. Unfortunately, pronunciations are scarce in the English Wiktionary (contrary to, e.g., the French one): only 44,795 articles provide transcriptions.

4.6. POS sections

4.6.1. Definitions, usage examples and linguistic labels

Within a POS section, definitions provide a sense inventory for a given word. A definition relates to a given meaning and contains a gloss and possibly one or several usage examples. Just as for etymologies, glosses and examples are available under four different formats: the original wikicode, an XML version where various elements are enclosed between markups,[4] a raw text version (more or less the XML version from which tags have been removed) and a syntactic parsing of the text in CoNLL format, performed by the Talismane parser (Urieli, 2013).
In entries describing highly polysemous words (*free, head, form, product, etc.*), definitions can display nested meanings. ENGLAWI's `level` attribute is used, as in Figure 4, to encode such nesting.[5] Definitions may include linguistic labels that signal attitudinal, diatopic, diachronic, diafrequential, diatechnical, diasemantic or grammatical information. For example, the Figure 4 illustrates that one of the many senses (to steal money) of the verb *strike* is an obsolete (diachronic categorization) intransitive (grammatical information) slang word (attitudinal categorization) that was used in British English (diatopic categorization).

```
<definition level="1">
  <gloss> <!-- ... --> <txt>To have a sharp or
    severe effect.</txt> <!-- ... --> </gloss>
    <!-- ... -->
    <definition level="2"><gloss>
      <labels>
        <label type="gram" value="intransitive"/>
        <label type="diatopic" value="UK"/>
        <label type="diachronic" value="obsolete"/>
        <label type="attitudinal" value="slang"/>
      </labels>
      <wiki>{{lb|en|intransitive|UK|obsolete|slang}}
            To [[steal]] [[money]].</wiki>
      <xml>To <innerLink>steal</innerLink>
           <innerLink>money</innerLink>.</xml>
      <txt>To steal money.</txt>
      <parsed>
      1 To    to    TO                0     _
      2 steal steal VB    t=inf       1     IM
      3 money money NN    n=singular  2     OBJ
      4 .     .     P                 1     SUB
      </parsed>
    </gloss>
  </definition> <!-- ... -->
</definition>
```

Figure 4: Labels in a nested definition of the verb *strike*

As can be seen in the wikicode shown in Figure 4, no category is given for each label. Labels of different types are indeed mixed up in a unique template: `{{lb|en|intransitive|UK|obsolete|slang}}`. We inventoried more than 1,100 labels that we manually classified in the above-mentioned categories. Our categories are inspired by (Hausmann, 1977) and (Hausmann et al., 1989) that we modified, according to current lexicographical practices. We also homogenized the values of similar labels (e.g. Wiktionary's *uncommon*, *rare* and *rare term* are all converted to *rare* in ENGLAWI; *comptheory* and *computing theory* have been homogenized to *computing theory*, etc.). Filtering ENGLAWI's definitions by attitudinal

[4] A comprehensive description of such markups is given in ENGLAWI's online documentation.

[5] Other illustrations are given in the documentation.

labels enables for example to build sentiment lexicons. Filtering them by diatopic labels allows to build lexicons of regional variants such as DIVAE (cf. Section 5.3.). Another type of linguistic label found in ENGLAWI's definitions provides information on selectional restrictions. For example, the general meaning of *boiling* is "that boils". This is also an informal hyperbole meaning "extremely hot or active" when applied to things and "feeling uncomfortably hot" when applied to persons (cf. Fig. 5).

```
<definition level="1"><gloss>
  <txt>That boils or boil.</txt>
</gloss></definition>
<definition level="1"><gloss>
  <labels>
    <label type="of" value="of a thing"/>
    <label type="attitudinal" value="informal"/>
    <label type="sem" value="hyperbolic"/>
  </labels>
  <txt>Extremely hot or active.</txt>
</gloss></definition>
<definition level="1"><gloss>
  <labels>
    <label type="of" value="of a person"/>
    <label type="attitudinal" value="informal"/>
    <label type="sem" value="hyperbolic"/>
  </labels>
  <txt>Feeling uncomfortably hot.</txt>
</gloss></definition>
```

Figure 5: Selectional restrictions for *boiling*

Wiktionary's definitions often include usage examples, sometimes coined or, most of the time, taken from diverse source. An illustration of ENGLAWI's corresponding content is depicted in Figure 6.

```
<exampleRef>
  <wiki>{{quote-book|year=2011|author=Divina Frau-Meigs|
    title=Media Matters in the Cultural Contradictions of
    the &amp;quot;Information Society&amp;quot;|page=299|
    passage=Issues such as verifiability (for age
    declared) anonymity (in spite of '''pseudos''' and
    avatars) and traceability are at stake[...]}}</wiki>
  <xml><quotation type="book">
   <attr type="year">2011</attr>
   <attr type="author">Divina Frau-Meigs</attr>
   <attr type="title">Media Matters in the Cultural
      Contradictions of the "Information Society"</attr>
   <attr type="pages">299</attr>
   <attr type="passage">Issues such as verifiability (for
      age declared), anonymity (in spite of pseudos and
      avatars) and traceability are at stake[...]</attr>
  </quotation></xml>
  <txt>Issues such as verifiability (for age declared),
      anonymity (in spite of pseudos and avatars) and
      traceability are at stake[...]</txt>
</exampleRef>
```

Figure 6: Quotation used as a usage example in the definition of the noun *pseudo* (sense #4/5 – Internet)

Definitions, together with etymologies, are the elements involving the most diverse wikicode templates: they occur not only in linguistic labels, but also in the glosses and examples. Moreover they are used to encode core content, not only style. Correct data extraction therefore requires consequent efforts to handle template properly. Such templates are generally not supported by coarse extractors and purely removed. As a consequence, 9% of DBnary's definitions (68,524 out of 760,184) are empty (they only contain a linguistic label, a dot or a curly bracket). For instance, DBnary's definition of *children* is a dot while ENGLAWI's text definition is *"plural of child."*, and the corresponding XML is: `<inflectionOf><inflectionType>plural</inflectionType> <lemma>child</lemma></inflectionOf>`
In addition to senses having empty definitions, the loss of senses for some entries (i.e. Wiktionary's senses that are absent from DBnary, even when the corresponding entry is present) is another effect of unsatisfying extraction. For example, the first four senses of the noun *pseudo* are present in DBnary, but the last sense is missing, probably due to a disregard of the `{{clipping of|en|pseudoephedrine}}` wiki template, which results in `<formOf type="clipping">pseudoephedrine</formOf>` in ENGLAWI's XML and *clipping of pseudoephedrine* in ENGLAWI's text definition. Besides not reimplementing most templates, DBnary disregards nested definitions, inducing a loss of word senses. For instance, DBnary features "only" 18 senses (out of 43) for the verb *strike*. The top-level meaning *"To have a sharp or severe effect"* is present but the more specific one *"to steal money"* (Fig. 4) is absent from the resource.

4.6.2. Lexical relations

Translations, semantic and morphological relations occur within POS sections. For example, the synonym *French person* relates to all the definitions (senses) of the second noun section of *frog* (Fig. 1). Conversely, some semantic relations (synonymy, antonymy, hypernymy and hyponymy) corresponding to a particular meaning also appear in definitions. For instance, *preface* and *epilogue* are meronyms for all senses of *book* (Fig. 7) while *tome* and *volume* are synonyms of only a given one (major division of a long work). Morphological sections contain derivation relations (cf. Fig. 8) and looser relations labeled *related*. Caution should be exercised when relying on this distinction: we observed that, in reality, words signaled by derivation relations in Wiktionary are often no real derivatives (morphological relations found in etymological sections are more trustworthy). Translations are given for national and regional languages, living and dead languages, natural and constructed languages (cf. Fig. 8).

```
<pos type="noun" lemma="1" etymNb="1">
  <definitions>
    <definition level="1">[...] A collection of
    sheets of paper [...]</definition>
    <definition level="1">[...] A major division of a
     long work <semRel type="syn">tome, volume</semRel>
     [...] </definition>
    <!-- ... -->
  </definitions>
  <section type="semRel">
    <item type="meronym">preface</item>
    <item type="meronym">epilogue</item>
  </section>
</pos>
```

Figure 7: Semantic relations for *book* (excerpt)

4.6.3. Inflectional paradigms

As seen above, both lemmas and inflections may appear as headwords in Wiktionary. When a headword relates to a lemma, the corresponding inflections (plural of nouns, comparative and superlative forms of adjectives and adverbs, participles and third-person forms of verbs, etc.) may be given below or next to the headword line. For ex-

```
<section type="morpho">
  <item type="derived">wrength</item>
  <item type="derived">wrongful</item>
  <item type="derived">wrongly</item>
</section>
<translations>
  <trans lang="ase">Y@Chin-PalmBack</trans>
  <trans lang="ca">incorrecte</trans>
  <trans lang="ca">erroni</trans>
  <trans lang="fr">erroné</trans>
  <trans lang="fr">incorrect</trans>
  <trans lang="io">vidar</trans>
  <trans lang="la">erroneus</trans>
  <trans lang="no">galt</trans>
  <trans lang="no">uriktig</trans>
  <trans lang="nn">feil</trans>
</translations>
```

Figure 8: Derivation relations and translations (American sign language, Catalan, French, Esperanto, Latin, Norwegian Bokmål, Norwegian Nynorsk) for *wrong* (excerpt)

ample, the inflections of the irregular verb *deal* are given in Wiktionary's headword line as depicted in Figure 9. When a headword relates to an inflected form, its definition usually provides the corresponding lemma and inflection type. An illustration is given in Figure 10 for the article *dealing*. Both kinds of information (either redundant or complementary) enable the generation of inflectional paradigms such as the verbal paradigm represented in Figure 11 for *deal*. These paradigms are directly used to produce the inflectional lexicon ENGLAFF (cf. Section 5.2.).

Verb
deal (*third-person singular simple present* **deals**, *present participle* **dealing**, *simple past and past participle* **dealt**)

Figure 9: Lemma's inflections in Wiktionary's headword lines: *deal*

Verb
dealing

1. *present participle of* **deal**

Figure 10: Plain text definition of an inflected form in Wiktionary: *dealing*

```
<paradigm>
  <inflection gracePOS="Vmn----" form="deal"/>
  <inflection gracePOS="Vm-ps--" form="dealt"/>
  <inflection gracePOS="Vmpp---" form="dealing"/>
  <inflection gracePOS="Vmip1s" form="deal"/>
  <inflection gracePOS="Vmip2s" form="deal"/>
  <inflection gracePOS="Vmip3s" form="deals"/>
  <inflection gracePOS="Vmip1p" form="deal"/>
  <inflection gracePOS="Vmip2p" form="deal"/>
  <inflection gracePOS="Vmip3p" form="deal"/>
  <inflection gracePOS="Vmis1s" form="dealt"/>
  <inflection gracePOS="Vmis2s" form="dealt"/>
  <inflection gracePOS="Vmis3s" form="dealt"/>
  <inflection gracePOS="Vmis1p" form="dealt"/>
  <inflection gracePOS="Vmis2p" form="dealt"/>
  <inflection gracePOS="Vmis3p" form="dealt"/>
</paradigm>
```

Figure 11: Inflectional paradigm for the verb *deal*

5. ENGLAWI's companions

5.1. G-PeTo

G-PeTo (GLAWI Perl Tools) is a set of scripts helping extract specific information from GLAWI dictionaries. These scripts can be used as is or they may be adapted to fit one's needs. They allow, for example, the extraction of headwords or whole articles matching specific criteria. Extracted articles are intended to be transformed with an XSL sheet, manually browsed or further queried by any other program. The scripts also enable the extraction of definitions including a given word or including a specific linguistic label. G-PeTo also includes the scripts used to generate ENGLAFF (cf. Section 5.2.) and DIVAE (cf. Section 5.3.).

5.2. ENGLAFF: a large inflectional lexicon of English

ENGLAFF is an inflectional lexicon containing 1,229,204 entries, each including an inflected form, its lemma and a morphosyntactic tag in GRACE format (Rajman et al., 1997). The number of inflections is given for each syntactic category in the last column of Table 1 (Section 4.2.). An excerpt of ENGLAFF is given in Figure 12.

```
bad|Afp--|bad
worse|Afc--|bad
worst|Afs--|bad
child|Nc-s|child
children|Nc-p|child
go|Nc-s|go
goes|Nc-p|go
go|Vmn----|go
go|Vmip1p|go
go|Vmip1s|go
go|Vmip2p|go
go|Vmip2s|go
go|Vmip3p|go
goes|Vmip3s|go
going|Vmpp---|go
gone|Vm-ps--|go
went|Vmis1p|go
went|Vmis1s|go
went|Vmis2p|go
went|Vmis2s|go
went|Vmis3p|go
went|Vmis3s|go
```

Figure 12: Excerpt of ENGLAFF

5.3. DIVAE: Diatopic Variation of English

DIVAE is a lexicon including 29,280 entries (19,172 distinct words) marked by 87 diatopic labels, extracted from ENGLAWI. Each entry of this lexicon contains: a word, its part of speech, the name of a place (area or country) where the word or specific meaning is used and a gloss of the word's meaning in that place. As illustrated in Figure 13, the entries include words that only exist in a given area or words that have a specific meaning in a particular place. For example, *reekin* exists only in Geordie (dialect spoken by Geordies, people from Tyneside) while only the meaning "impertinent, assertive" of the Caribbean *mannish* is specific to that particular place. Besides linguistic studies on diatopic variation of English, an early version of DIVAE has proven useful for author profiling (Tanguy et al., 2011).

```
aubergine Nc UK          an Asian plant, Solanum
                         melongena, cultivated for [...]
chap      Nc Australia   A man, a fellow.
chap      Nc Scotland    A blow; a rap.
chap      Nc Southern US A child.
chap      Nc UK          A man, a fellow.
chap      Nc UK          A customer, a buyer.
closet    Nc Ireland     Any small room or side-room [...]
closet    Nc Scotland    Any small room or side-room [...]
closet    Nc Scotland    A sewer.
closet    Nc UK          clipping of closet of ease [...]
closet    Nc US          One used to store food [...]
closet    Nc US          One intended for storing clothes
eggplant  Nc Australia   The plant Solanum melongena.
eggplant  Nc New Zealand The plant Solanum melongena.
eggplant  Nc US          The plant Solanum melongena.
mannish   A  Caribean    Impertinent, assertive.
reekin    A  Geordie     Totally stinking.
```

Figure 13: DIVAE: Examples of English diatopic variants

5.4. WIND: Wiktionary's Inclusion Dates

This lexicon contains the inclusion dates of Wiktionary's headwords in its nomenclature. WIND has been created by parsing Wiktionary's *history dump*, which contains every version of all articles (stored after each individual contributor's edition). Each entry of the resource has four fields: a written form, the inclusion date of this written form (i.e. the creation date of the corresponding page), a part of speech and the creation date of the section corresponding to that POS. The first entries (*abaca, abacinate*) have been imported from the 1913 edition of the *Webster's New International Dictionary of the English Language* when Wiktionary was created, in December 2002. POS sections were added later (cf. Fig 14). Recent entries (*protomyth, mother-hen, subflow*) are generally created nowadays together with their POS section(s). This resource is useful for the study of Wiktionary's evolution. It is also relevant to neology studies and metalexicographic descriptions (Sajous et al., 2018a).

Entry	Entry_inclusion	POS	POS_inclusion
abaca	2002-12-15	NC	2003-03-17
abacinate	2002-12-15	V	2003-02-05
...			
free	2002-12-12	ADJ	2003-10-20
free	2002-12-12	V	2003-10-20
free	2002-12-12	ADV	2005-10-30
free	2002-12-12	NC	2007-05-10
...			
protomyth	2019-08-25	NC	2019-08-25
mother-hen	2019-08-25	V	2019-08-25
subflow	2019-08-25	NC	2019-08-25
subfibril	2019-08-25	NC	2019-08-25

Figure 14: WIND: Inclusion dates of headwords

6. Using GLAWI dictionaries

We enumerated in Section 2 works that use data extracted from Wiktionary. We recap below how GLAWI, GLAWIT and ENGLAWI have proven useful for NLP and linguistics. Then, we describe a new ongoing work on computing word embeddings from ENGLAWI's definitions.

6.1. NLP and linguistic studies

Since their creation, GLAWI dictionaries have been used for various NLP works: Navarro et al. (2009) and Sajous et al. (2013b) used WiktionaryX, the ancestor of GLAWI and ENGLAWI to tune random walks algorithms intended to improve the synonymy networks extracted from this resource. The diatopic variants of English words included in WiktionaryX have been used by Tanguy et al. (2011) for authorship attribution. Hathout et al. (2014b) used GLAWI to create GLÀFF, the largest inflectional and phonological lexicon available for French (the French equivalent of ENGLAFF). Hathout et al. (2014a) used GLAWI's definitions and morphological relations explicitly present in the dictionary to learn new morphological relations. Calderone et al. (2017) used GLAWIT to design a method for Italian stress prediction. Pierrejean and Tanguy (2018) used ENGLAWI to assess the influence of the degree of polysemy of words on the variability of word embeddings.

NLP is thus a natural MRD consumer. Corpus linguistics and metalexicographical studies also benefit from GLAWI dictionaries. Flaux et al. (2014) have collected human names that denote a creative activity (e.g. *symphoniste* 'symphonist', *sculpteur* 'sculptor', *romancier* 'novelist', etc.) from professional dictionaries and from web harvesting. A simple lookup in GLAWI's glosses, based on lexical cues only (i.e. looking for already inventoried names in Aristotelian definitions), enabled a 15% increase of the initial database. In a study on motion verbs and evaluative morphology, Stosic and Amiot (2019) extracted evaluative verbs from the *Trésor de la Langue Française*, a large academic dictionary. According to the authors, one limitation of their approach stems from using this dictionary that has not been updated since its initial publication and which only records words that are really lexicalized. Looking for specific affixes in GLAWI's nomenclature enabled the authors to expand the initial list of French evaluative verbs from 171 up to 940. On a similar line, Stosic (2019) extended a database of motion verbs (from 521 to 960 verbs) by exploiting GLAWI's genus-differentia definitions. Because Wiktionary records neologisms earlier that other dictionaries, and sometimes provide culturally-informed features in semantic description that is not found in professional dictionaries, WIND can been used for real-time neology watch both for general language (Sajous et al., 2018a) and for specialized vocabulary (Sajous et al., 2018b).

6.2. Lexicographic word embeddings

Distributional semantic models (DSM) are used to identify semantic neighbors of words. Among them, the (neural) word embeddings are usually seen as a way to provide compact and fixed-size representations of word meaning. In the last decade, they have become a central component in most NLP systems. What is their real descriptive and explanatory adequacy is still an open question addressed by Lenci (2018). For instance, he argues that DSM tend to identify semantic relatedness rather than semantic similarity and that their ability to properly distinguish different semantic relations is limited. DSM are generally trained on large corpora and evaluated on standard data sets. Most studies usually compare different architectures of some systems and assess their respective performances. The other factor that we can change is the training data: one could compare different systems trained on different inputs. Instead of training the DSM on large corpora, one could use lexicographic definitions as input. DSM trained on definitions have been proposed by Noraset et al. (2016) and Bosc and Vincent (2018). In spite of the reduced size of the data, the performance of these models are comparable to the performance of models trained on (very) large corpora. These "lexicographic word embeddings" can be created by standard tools such as Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017) or LSTM architectures (Hochreiter and Schmidhuber, 1997). In the following, we compare 3 models: a classical FastText model trained on the Common Crawl, a 600 billion words corpus (Grave et al., 2018), a FastText model trained on ENGLAWI's definitions and a model created by means of an LSTM, also trained on ENGLAWI's definitions.

The words that occur in corpora or data sets are sometimes POS tagged but almost never sense tagged. In other words, several definitions may correspond to a polysemous word that has not been disambiguated. For this reason, we con-

catenated all the definitions of a polysemous entry in order to obtain an "average meaning" that includes all its acceptions. The concatenated definitions are used both to train the lexicographic FastText model and as input of the LSTM. At first sight, a corpus of definitions is unsuitable to train distributional models such as Word2Vec or FastText, because the definitions do not contain the words they define. In a corpus made up of (concatenated) definitions, the entries do not have their definitions as contexts and only the words used in the definitions will have representations. In order to alleviate this problem, we adopted the solution proposed by Noraset et al. (2016) and Bosc and Vincent (2018) which is to use a corpus where each headword is followed by the concatenation of all its definitions. In this way, headwords are next to the most informative part of the definitions, their first words. We trained two word embeddings on this corpus, one with FastText, as we said before, and one with Word2Vec to code the input of the LSTM. The FastText model was computed with skipgram and the default values for all the parameters. The Word2Vec model was computed with skipgram and a minimum window size of 15. The LSTM was trained with the headwords as targets and the concatenations of definitions as input, where each word is represented by its Word2Vec embedding. After training, for each definition, the last hidden activation of the LSTM represents the embedding for that definition. Table 3 reports, for the three models and inputs, the first 10 neighbors of the word *godhood*, defined in ENGLAWI as *the state of being a god; divinity*. In this example, we see that the three models produce different kinds of semantic neighborhood. FastText, trained on the Common Crawl corpus, produces neighbors whose connections to the target word correspond to various types of relations: attributes (e.g. *godhead*), property or quality (e.g. *omnipotence*), process (e.g. *deification*), etc. In the two models trained on ENGLAWI, the neighbors denote states (e.g. *nirvana*) or conditions (e.g. *fatherhood*). However, we observe a difference between the two. Some neighbors in the lexicographic FastText model are not semantically associated with *godhood* (e.g. *fatherhood*, *motherhood*, *selfhood*, *manhood*, etc.) and denote other conditions. Others, such as *triune*, *nirvana* and *incarnation*, are semantically "related" rather than truly "similar". On the other hand, the neighbors in the LSTM model feature a stronger semantic similarity with the target word (*deityhood*, *blessedness*, *divineness*, *angelhood*, etc.). A possible explanation is that LSTMs are able to recognize frequent patterns in the definitions. This capability is further enhanced by the use of the Word2Vec vectors as inputs. For instance, in the Word2Vec model, words like *state, quality* and *condition* tend to be represented by close vectors. In this way, the LSTM is able to generalize over the lexical variations as for the following definitions:

(a) *state of being a deity; divinity* (deityhood)

(b) *the quality of being divine* (divineness)

(c) *the state or condition of being blessed, holy* (blessedness)

The single *godhood* example, given in Table 3, is telling. However, it cannot, alone, definitely answer to the kind of questions raised by Lenci (2018). Further qualitative and quantitative analysis are indeed required to explore the nature of the semantic neighborhood captured by the different models, both from an NLP and a linguistic perspective.

FastText Common Crawl	FastText ENGLAWI	LSTM ENGLAWI
godhead	fatherhood	deityhood
deification	demigod	blessedness
godlike	motherhood	divineness
immortality	selfhood	paganity
deific	triune	angelhood
divinity	childhood	deathlessness
omnipotence	nirvana	fathership
divinization	bodhisattva	worshipability
deity	manhood	creatorship
demigod	incarnation	buddhahood

Table 3: Ten first neighbors of *godhood*

7. Conclusion

In this paper, we presented ENGLAWI, a machine-readable dictionary extracted from Wiktionary. Unlike other approaches that aim to extract massively multilingual data, we implemented a fine extraction of the English language edition of Wiktionary, that enabled us to extract the full body of information available in that dictionary, for that language, and to encode it in a structured and normalized MRD. Regarding the size of the extracted nomenclature, as well as that of the inflectional paradigms, we achieve better results than previous existing works. We have also shown that our approach to information extraction is more reliable and produces data that is consistent with Wiktionary's content. A strength of ENGLAWI, among all lexical information it includes, lies in the fact that it is the only resource that 1) contains definitions of contemporary English words, including recent neologisms 2) is free and 3) features a large coverage. The number of works in NLP and linguistics that successfully relied on these definitions and on the morphological information included in ENGLAWI and other GLAWI dictionaries may be a positive answer to the question we raised at the beginning of the paper: Yes, MRD are useful for NLP and linguistics. As for the question asked by Ide and Veronis (1993) about the extraction of knowledge from MRD (*Have we wasted our time ?*), we may give a negative answer: even if the methods designed in the 1990s are not to be reimplemented (different times, different data, methods and computational capabilities), we consider the researchers who worked on extracting knowledge from MRD by that time paved the way to more recent studies. We have shown in a preliminary study that the lexicographic word embeddings computed from ENGLAWI's definitions produce different results than those obtained from corpora. This first result encourages us to continue this work: in an upcoming study, we plan to compare more systematically the different types of similarity captured by the word embeddings trained on dictionaries on one side and those trained on corpora on the other.

8. Acknowledgements

Syntactic parsing has been performed using the OSIRIM platform, that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government and ERDF.

9. Bibliographical References

Ács, J. (2014). Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1938–1942, Reykjavik.

Bański, P., Bowers, J., and Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In *Proceedings of ELex 2017 Conference*, Leiden.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bosc, T. and Vincent, P. (2018). Auto-Encoding Dictionary Definitions into Consistent Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels. Association for Computational Linguistics.

Calderone, B., Sajous, F., and Hathout, N. (2016). GLAWIT: A free large Italian dictionary encoded in a fine-grained XML format. In *Proceedings of the 49th Annual Meeting of the Societas Linguistica Europaea (SLE 2016)*, pages 43–45, Naples.

Calderone, B., Pascoli, M., Sajous, F., and Hathout, N. (2017). Hybrid Method for Stress Prediction Applied to GLAFF-IT, a Large-scale Italian Lexicon. In Jorge Gracia, et al., editors, *Language, Data, and Knowledge*, pages 26–41, Cham. Springer International Publishing.

Creese, S., McGillivray, B., Nesi, H., Rundell, M., and Sule, K. (2018). Everything You Always Wanted to Know about Dictionaries (But Were Afraid to Ask): A Massive Online Course. In *Proceedings of the 18th EURALEX International Congress*, pages 59–66, Ljubljana.

De Smedt, T., Marfia, F., Matteucci, M., and Daelemans, W. (2014). Using Wiktionary to Build an Italian Part-of-Speech Tagger. In Elisabeth Métais, et al., editors, *Proceedins of the 19th International Conference on Application of Natural Language to Information Systems,* Lecture Notes in Computer Science, volume 8455. Springer, Cham.

Flaux, N., Lagae, V., and Stosic, D. (2014). Romancier, symphoniste, sculpteur : les noms d'humains créateurs d'objets idéaux. In *Actes du 4eme Congrès Mondial de Linguistique Française (CMLF 2014)*, pages 3075–3089, Berlin.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, Miyazaki.

Gurevych, I., Eckle-Kohler, J., and Matuschek, M. (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Morgan & Claypool Publishers.

Hathout, N. and Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5):125–168.

Hathout, N. and Sajous, F. (2016). Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož.

Hathout, N., Sajous, F., and Calderone, B. (2014a). Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary. In *Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, pages 65–74, Dublin.

Hathout, N., Sajous, F., and Calderone, B. (2014b). GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik.

Hausmann, F. J., Wiegand, O., Zgusta, H., and Zgusta, L. (1989). *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*. Walter de Gruyter, New-York.

Hausmann, F. J. (1977). *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Max Niemeyer Verlag, Tübingen.

Hellmann, S., Brekle, J., and Auer, S. (2013). Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud. In Hideaki Takeda, et al., editors, *Semantic Technology*, volume 7774 of *Lecture Notes in Computer Science*, pages 191–206. Springer Berlin Heidelberg.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ide, N. and Veronis, J. (1993). Extracting Knowledge Bases from Machine-Readable Dictionaries: Have We Wasted Our Time. In *Proceedings of the KB & KS'93 workshop*, pages 257–266, Tokyo.

Ide, N. and Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, 29(2):167–179.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126, Portorož.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Liebeck, M. and Conrad, S. (2015). IWNLP: Inverse Wiktionary for Natural Language Processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 414–418, Beijing.

Metheniti, E. and Neumann, G. (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)*, pages 147–161, Oslo.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information*

Processing Systems (NIPS'13) - Volume 2, pages 3111–3119, Lake Tahoe.

Nastase, V. and Strapparava, C. (2015). knoWitiary: A Machine Readable Incarnation of Wiktionary. *International Journal of Computational Linguistics and Applications*, 6(2):61–82.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., and Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27, Singapore.

Navigli, R. and 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*, pages 216–225, Uppsala.

Noraset, T., Liang, C., Birnbaum, L., and Downey, D. (2016). Definition Modeling: Learning to define word embeddings in natural language. *CoRR*, abs/1612.00394.

Pierrejean, B. and Tanguy, L. (2018). Predicting Word Embeddings Variability. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 154–159, New Orleans.

Rajman, M., Lecomte, J., and Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., George, M., Pet, M., and Bański, P. (2019). LMF Reloaded. In *Proceedings of the 13th Conference of the Asian Association for Lexicography (AsiaLex 2019)*, pages 533–539, Istanbul.

Sajous, F. and Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the eLex 2015 conference*, pages 405–426, Herstmonceux, England.

Sajous, F., Hathout, N., and Calderone, B. (2013a). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 285–298, Les Sables d'Olonne, France.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2013b). Semi-automatic enrichment of crowdsourced synonymy networks: the WISIGOTH system applied to Wiktionary. *Language Resources and Evaluation*, 47(1):63–96.

Sajous, F., Josselin-Leray, A., and Hathout, N. (2018a). The Complementarity of Crowdsourced Dictionaries and Professional Dictionaries viewed through the Filter of Neology. *Lexis*, 12.

Sajous, F., Josselin-Leray, A., and Hathout, N. (2018b). Définir la néologie terminologique dans les dictionnaires généraux : le domaine de l'informatique analysé par « les foules » et par les professionnels... de la lexicographie. In *4ème Congrès international de néologie des langues romanes (CINEO 2018)*, Lyon.

Schlippe, T., Ochs, S., and Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH'2010)*, pages 2290–2293, Makuhari, Chiba.

Segonne, V., Candito, M., and Crabbé, B. (2019). Using Wiktionary as a resource for WSD: the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics*, pages 259–270, Gothenburg.

Sérasset, G. (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul.

Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web*, 6(4):355–361.

Stosic, D. and Amiot, D. (2019). Motion verbs and evaluative morphology. In Michel Aurnague et al., editors, *The semantics of dynamic space in French: descriptive, experimental and formal studies on motion expression*, pages 179–216. John Benjamins, Amsterdam/Philadelphia.

Stosic, D. (2019). Manner as a cluster concept: What does lexical coding of manner of motion tell us about manner? In Michel Aurnague et al., editors, *The semantics of dynamic space in French: descriptive, experimental and formal studies on motion expression*, pages 142–177. John Benjamins, Amsterdam/Philadelphia.

Tanguy, L., Urieli, A., Calderone, B., Hathout, N., and Sajous, F. (2011). A multitude of linguistically-rich features for authorship attribution. In *Notebook for PAN at CLEF 2011*, Amsterdam.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II-Le Mirail.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech.