Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary

Nabil Hathout, Franck Sajous

Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary

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

Abstract

GLAWI is a free, large-scale and versatile Machine-Readable Dictionary (MRD) that has been extracted from the French language edition of Wiktionary, called Wiktionnaire. In (Sajous and Hathout, 2015), we introduced GLAWI, gave the rationale behind the creation of this lexicographic resource and described the extraction process, focusing on the conversion and standardization of the heterogeneous data provided by this collaborative dictionary. In the current article, we describe the content of GLAWI and illustrate how it is structured. We also suggest various applications, ranging from linguistic studies, NLP applications to psycholinguistic experimentation. They all can take advantage of the diversity of the lexical knowledge available in GLAWI. Besides this diversity and extensive lexical coverage, GLAWI is also remarkable because it is the only free lexical resource of contemporary French that contains definitions. This unique material opens way to the renewal of MRD-based methods, notably the automated extraction and acquisition of semantic relations.

Keywords: French, Machine-Readable Dictionary, Free Lexical Resource, Wiktionary, Wiktionnaire

1. Introduction

GLAWI[1] is a large Machine-Readable Dictionary (MRD) extracted from Wiktionnaire, the French language edition of Wiktionary, and converted into a workable XML format. In a previous work, Sajous et al. (2010) introduced WiktionaryX, an electronic lexicon including lemmas, semantic relations and translations. Hathout et al. (2014b) described how GLÀFF, a large inflectional and phonological lexicon, has been extracted from the same source. The assessment of GLÀFF's lexical coverage and the quality of its phonemic transcriptions has shown that Wiktionnaire is a valuable starting point to build lexical resources of good quality. Sajous and Hathout (2015) introduced GLAWI, a dictionary built from an updated version of Wiktionnaire that merges the information stored in WiktionaryX and GLÀFF into a single resource. New information, such as etymology and morphological relations, has also been added. Sajous and Hathout (2015) focused on the parsing process and the standardization of Wiktionnaire's heterogeneous data, a prerequisite to produce a workable MRD. In the current article, we illustrate the richness of GLAWI's lexical knowledge, leaving apart the extraction process. We also contemplate different uses that can be made of this resource, either in academic research, or in concrete NLP applications.

2. Resource description

The general structure of GLAWI's entries is illustrated in Figure 1. GLAWI's macro- and micro-structure are very close to the ones of Wiktionnaire: the basic unit is a word written form (hereafter, grapheme), associated with a given page/URL. When several parts of speech (POS) or homographs correspond to the same grapheme, the article contains one separate POS section for each one of them. Each POS section includes definitions (glosses and examples), and several optional subsections described hereafter. Table 1 gives the number of lemmas and inflected forms by POS.

[1]GLAWI is freely available at `http://redac.univ-tlse2.fr/lexicons/glawi_en.html`

POS	Lemmas	Inflected forms
noun	179,340	272,170
proper noun	57,371	8,019
adjective	56,296	93,295
verb	36,928	1,251,809
adverb	5,552	5,552
total	335,487	1,630,845

Table 1: Lemmas and inflected forms for the main POS

2.1. Definitions

Word senses, marked by **`definition`** tags, are listed in the POS sections. A definition contains a gloss and possibly one or several usage examples. Glosses and examples are each available in four different versions (an example is given in Figure 2):

1. the original wikicode, intended for developers willing to perform specific extractions or conversions.

2. an XML formatted version where markups encode typesetting (boldface, italic, etc.), dates, foreign words, mathematical/chemical formulae and external/inner links. Markups can be used to select or to remove specific types of elements (e.g. foreign words or non textual content such as formulae). Links could be used by a weighting scheme in information retrieval (Cutler et al., 1997) or to build hyperlink graphs for semantic similarity computation (Weale et al., 2009).

3. a raw text version. Many other text versions can be generated from the XML one by selecting specific elements and formatting them differently.

4. a CoNLL output (Nivre et al., 2007) of the Talismane syntactic parser (Urieli, 2013). Dependencies may prove useful for various tasks. For example, Hathout et al. (2014a) used them as features to train a classifier and identify Wiktionnaire's glosses of derived action nouns, with an accuracy ranging from 94% to 99%.

```
<article>
  <title>mousse</title>
  <pageId>7930</pageId>
  <meta>
    <category>Lexique en français de la navigation</category>
    <category>Noms multigenres en français</category>
    <reference>TLFi</reference>
  </meta>
  <text>
    <pronunciations>
      <pron region="France">mus</pron>
    </pronunciations>
    <pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
      <pronunciations>
        <pron>mus</pron>
      </pronunciations>
      <paradigm>
        <wiki>{{fr-rég|mus}}</wiki>
        <inflection form="mousse" gracePOS="Ncfs" pron="mus"/>
        <inflection form="mousses" gracePOS="Ncfp" pron="mus"/>
      </paradigm>
      ...
    </pos>
    <pos type="nom" lemma="1" locution="0" homoNb="2" gender="m" number="s">
    ...
    </pos>
    <pos type="nom" lemma="1" locution="0" homoNb="3" gender="m" number="s">
    ...
    </pos>
    <pos type="adjectif" lemma="1" locution="0" gender="e" number="s">
    ...
    </pos>
    <pos type="verbe" lemma="0" locution="0">
      <inflectionInfos>
        <inflected gracePOS="Vmip1s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmip3s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmsp1s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmsp3s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmmp2s-" lemma="mousser" pron="mus"/>
      </inflectionInfos>
    </pos>
  </text>
</article>
```

Figure 1: General structure of an article in GLAWI: *mousse* entries

2.2. Labels

As shown in Figure 2, definitions may include linguistic labels. They are identified by the parser and marked with **`label`** tags. Moreover, we inventoried thousands of labels and manually assigned to each one a category among the followings: *attitudinal, diachronic, diafrequential, diatopic, domain, grammar, loan, semantics* or *other* for uninventoried labels.[2] GLAWI's main linguistic labels are listed in Table 2. They can be used to study lexical variation. They may also prove useful for various applications. words marked as attitudinal may be used for sentiment analysis. Specialized lexicons can be extracted on the basis of domain labels. Words marked with these labels can also be used as seeds for focused web-crawling. Diafrequential labels may guide text simplification by favoring *more usual* words rather than *very rare* ones. Diatopic and diachronic labels may be leveraged in text classification, for instance, when building a corpus from the Web. Texts featuring a large number of *dated* or *archaic* words are likely to be archived historical documents. GLAWI's diatopic variations may help distinguish closely related languages, for example hexagonal and overseas French. Blacklisted words based on such labels could be used to improve state-of-the-art classifiers, as Tiedemann and Ljubešić (2012) did to discriminate between Bosnian, Croatian and Serbian. Such lexicons may reveal French or Canadian origin in author profiling or identification, in a similar way to Tanguy et al. (2011), who used British/American English variants as features for author attribution.

[2] More details are given in (Sajous and Hathout, 2015).

```
<definition>
  <gloss>
    <labels>
      <label type="sem" value="métonymie"/>
      <label type="attitudinal" value="familier"/>
    </labels>
    <wiki>{{méton|fr}} {{familier|fr}} [[bière|Bière]]</wiki>
    <xml><innerLink ref="bière">Bière</innerLink></xml>
    <txt>Bière</txt>
    <parsed>1 Bière bière NC nc  g=f|n=s    0  root 0   root  100,00 55,43 98,84</parsed>
  </gloss>

  <example>
    <wiki>'' Une bonne '''mousse''' bien fraîche, sans faux-col est un oxymore.''</wiki>
    <xml><i> Une bonne <b>mousse</b> bien fraîche, sans faux-col est un oxymore.</i></xml>
    <txt> Une bonne mousse bien fraîche, sans faux-col est un oxymore.</txt>
    <parsed>1 "     "     PONCT PONCT _     11    ponct 11    ponct 100,00 79,93 99,76
          2  Une   une   DET   DET   g=f|n=s 4   det   4     det   100,00 98,84 99,67
          3  bonne bon   ADJ   adj   g=f|n=s 4   mod   4     mod   100,00 98,18 98,91
          4  mousse mousse NC  nc    n=s   11    suj   11    suj   100,00 91,02 89,73
          5  bien  bien  ADV   adv   _     6     mod   6     mod   100,00 85,58 86,14
          6  fraîche frais ADJ adj   g=f|n=s 4   mod   4     mod   100,00 71,25 98,98
          7  ,     ,     PONCT PONCT _     11    ponct 11    ponct 100,00 92,74 98,89
          8  sans  sans  P     P     _     11    mod   11    mod   100,00 88,98 98,01
          9  faux-col    _     NC    _     _     8     prep  8     prep  100,00 65,45 81,85
          10 "     "     PONCT PONCT _     11    ponct 11    ponct 100,00 95,95 86,35
          11 est   être  V     v     n=s|p=3|t=pst 0   root  0     root  100,00 96,82 99,88
          12 un    un    DET   DET   g=m|n=s 13  det   13    det   100,00 83,94 99,52
          13 oxymore _   NC    _     _     11    ats   11    ats   100,00 60,64 77,93
          14 .     .     PONCT PONCT _     11    ponct 11    ponct 100,00 100,00 99,80
    </parsed>
  </example>
</definition>
```

Figure 2: A given sense of *mousse* (feminine noun, homograph #1) as a metonym for *bière* 'bier'

```
<etymology>
  <etym>
    <labels>
      <label type="diachronic" value="1759"/>
    </labels>
    <wiki>{{date|1759}} du {{étyl|grc|fr|μονόξυλοσ|monoxylos|}}
        {{cf|mono-|-xyle|lang=fr}}.</wiki>
    <xml><date>1759</date> du grec ancien
        <foreignWord lang="grc" translit="monoxylos">μονόξυλοσ</foreignWord>
        <cf value="mono-|-xyle" lang="fr"/>.</xml>
    <txt>du grec ancien μονόξυλοσ monoxylos voir mono- et -xyle.</txt>
    <parsed>1 du    de    P+D   P+D   g=m|n=s 5   mod   5     mod   100,00 49,16 86,08
2     grec  grec  NC    nc    g=m|n=s 1   prep  1     prep  100,00 37,48 96,64
3     ancien ancien ADJ adj   g=m|n=s 2   mod   2     mod   100,00 82,63 97,10
4     μονόξυλοσ   _     NPP   _     _     2     mod   2     mod   100,00 21,27 92,84
5     monoxylos   _     V     _     _     0     root  0     root  100,00 44,77 99,11
6     voir  voir  VINF  v     _     5     obj   5     obj   100,00 50,63 76,69
7     mono- _     ADV   _     _     6     mod   6     mod   100,00 19,45 75,96
8     et    et    CC    CC    _     6     coord 6     coord 100,00 30,70 76,45
9     -xyle _     NPP   _     _     8     dep_coord   8     dep_coord   38,96 54,48 96,89
10    .     .     PONCT PONCT _     5     ponct 5     ponct 100,00 100,00 99,07
</parsed>
  </etym>
</etymology>
```

Figure 3: Etymology of *monoxyle* 'dugout'

Diafrequential		**6,166**
rare	rare	4,215
extrêmement rare	extremely rare	1,016
très rare	very rare	301
plus courant	more common	190
courant	common	186
plus rare	more rare	176
moins courant	less common	62
peu usité	rarely used	20
Diachronic		**24,450**
vieilli	old	9,431
désuet	dated	6,043
avant 1835	before 1835	1,654
néologisme	neologism	820
archaïque	archaic	661
1986		73
1990		72
	766 other years	5,841
Loanwords		**1,493**
anglicisme	Anglicism	1,446
indo-européen commun	usual indo-european	22
hispanisme	Hispanism	11
germanisme	Germanism	7
gaulois	Gallic	4
catalan	Catalan	3
Semantics		**23,860**
figuré	figurative	10,859
par extension	by extension	6,666
en particulier	in particular	2,574
analogie	analogy	1,213
métonymie	metonymy	886
ellipse	ellipsis	793
spécialement	especially	704
métaphore	metaphor	75
hyperbole	hyperbole	30
apocope	apocope	24
généralement	generally	19
litote	litote	10
figure	rethorical figure	7

Diatopic		**8,726**
Québec	Quebec	1,717
France	France	1,138
Canada	Canada	971
Suisse	Switzerland	962
Belgique	Belgium	637
Lorraine	Lorraine	299
Occitanie	Occitanie	246
Normandie	Normandie	134
Provence	Provence	123
Acadie	Acadie	122
Louisiane	Louisiana	90
Réunion	Réunion	89
Afrique	Africa	64
Congo-Kinshasa	Congo-Kinshasa	47
Ardennes	Ardennes	46
Languedoc-Roussillon	Languedoc-Roussillon	44
Bretagne	Brittany	40
	362 other areas	1,957
Domains		**155,532**
localités	locality	49,060
géographie	geography	11,935
botanique	botanic	6,461
zoologie	zoology	5,460
médecine	medecine	5,258
chimie	chemistry	3,358
histoire	history	2,804
marine	sailing	2,644
religion	religion	2,559
linguistique	linguistics	2,177
agriculture	agriculture	2,071
anatomie	anatomy	2,005
informatique	computer science	1,718
droit	law	1,698
physique	physics	1,579
militaire	military	1,572
musique	music	1,570
minéralogie	mineralogy	1,531
biologie	biology	1,515
antiquité	antique	1,327
cuisine	cooking	1,284
	367 other domains	45,946

Table 2: Main linguistic labels used in definitions and etymology sections. Translations are given in the right column.

2.3. Etymology

85 % of the pages describing a lemma include an etymology section. Figure 3 shows the etymology for *monoxyle* 'dugout'. Etymologies are available in the four formats listed in Section 2.1.: original wikicode, XML, raw text and CoNLL versions. The information given in Figure 3 includes an attestation date (1759), a source language (Ancient Greek) and a morphological structure (*mono-*|*-xyle*). Indications about words formation may be used to complement the morphological relations (cf. section 2.6.). Optional words' transliterations may also be given when words are written in non Latin alphabets. For example, the transliteration *monoxylos* is provided for the Greek *μονόξλοσ*. The meaning of the etymon in the source language may also be given as an attribute. Figure 4 illustrates that the sense of the Romani etymon *piyav* of the French *pillaver*, is *boire* "drink alcohol'. The main languages of origin of the French words mentioned in the etymology sections are listed in Table 3.

```
<foreignWord lang="rom"
             sense="boire">piyav</foreignWord>
```

Figure 4: Meaning of the Romani *piyav*, found in the etymology of the French *pillaver*

# Etym	Language	Examples
17,093	Latin	*bibliothèque* 'library', *optimum* 'optimum'
5,954	Greek	*monoxyle* 'dugout', *pédagogie* 'pedagogy'
4,403	English	*self-service*, *syllabification*
2,935	Occitan	*resquiller* 'to queue-jump', *escalade* 'climbing'
1,732	Old French	*empoté* 'clumsy', *se débiner* 'to leave secretly'
1,189	Italian	*bambin* 'toddler', *mandoline* 'mandoline'
775	Spanish	*aficionado* 'fan', *sieste* 'nap'
712	Arabic	*algèbre* 'algebra', *baroud* 'combat'
591	German	*ersatz* 'inferior quality substitute', *nouille* 'noodle'
400	Japanese	*kanji* 'kanji', *kimono* 'kimono'
311	Russian	*chaman* 'shaman', *bélouga* 'beluga'
264	Frankish	*fauteuil* 'armchair', *hache* 'axe'
244	Catalan	*paella* 'paella', *salicorne* 'samphire'
207	Breton	*cohue* 'rabble', *menhir* 'menhir'
197	Dutch	*havre* 'harbor', *maquignon* 'horse trader'
196	Portuguese	*caravelle* 'caravel', *piranha* 'piranha'
175	Gaulish	*trogne* 'mug (face)', *andain* 'swath'
164	Hebrew	*talmud* 'Talmud', *schwa* 'schwa'
163	Basque	*jokari* 'Jokari', *axoa* (Basque veal stew)
138	Sanskrit	*nirvana* 'nirvana', *gourou* 'guru'
+ 3,155	etymologies in 306 other languages	
Total: 40,410	etymologies in 326 different languages	

Table 3: 20 most frequently mentioned languages in GLAWI's etymology sections

2.4. Semantic relations

POS sections may include (quasi-)synonyms/antonyms, hypernyms/hyponyms, meronyms/holonyms and troponyms. An example of such relations is given in Figure 5 for the noun *communisme* 'communism'. The number of semantics sections per POS and the total number of semantic relations are given in Table 4.

```
<subsection type="semRel">
  <item type="synonym">collectivisme</item>
  <item type="synonym">marxisme</item>
  <item type="antonym">capitalisme</item>
  <item type="hyperonym">idéologie</item>
  <item type="hyponym">bolchévisme</item>
  <item type="hyponym">léninisme</item>
  <item type="hyponym">trotskisme</item>
</subsection>
```

Figure 5: Semantic relations for *communisme*

Such lexical semantic links may prove useful for various applications such as lexical substitution (McCarthy and Navigli, 2009), metaphor resolution (Desalle et al., 2009) or when setting up protocols for the detection of pathologies (Desalle et al., 2014).

2.5. Translations

POS sections often include translations in various languages. Figure 6 gives an example of translations for *piste cyclable* 'bicycle path'. We can see that languages such as Norwegian Bokmål and Norwegian Nynorsk have two different language codes. The number of translations per POS is given in Table 5.

Semantic Relations			
POS	**# sections**	**Relations**	
nouns	31,332		
		synonym	46,605
		near-synonym	2,454
		antonym	4,625
		hyperonym	20,093
		hyponym	21,472
		holonym	1,115
		meronym	2,566
adjectives	5,613		
		synonym	8,854
		near-synonym	833
		antonym	3,858
		hyperonym	483
		hyponym	1,062
		holonym	23
		meronym	34
verbs	5,157		
		synonym	9,322
		near-synonym	643
		antonym	1,675
		hyperonym	86
		hyponym	162
		troponym	125
adverbs	1,491		
		synonym	3,041
		near-synonym	196
		antonym	494

Table 4: Semantic relations

```
<translations>
  <trans lang="de">Radweg</trans>
  <trans lang="en">bicycle path</trans>
  <trans lang="it">pista ciclabile</trans>
  <trans lang="it">ciclopista</trans>
  <trans lang="nl">fietspad</trans>
  <trans lang="no_nb">sykkelvei</trans>
  <trans lang="no_nn">sykkelveg</trans>
  <trans lang="pt">ciclovia</trans>
  <trans lang="sv">cykelväg</trans>
</translations>
```

Figure 6: Translations for *piste cyclable* 'bicycle path'

Translations		
POS	**# sections**	**# translations**
nouns	71,133	383,612
adjectives	16,797	60,360
verbs	11,484	70,615
adverbs	3,014	14,478
total	102,428	529,065

Table 5: Translations

Many applications may benefit from these translations. Statistical machine translation algorithms tend to disregard lexicons. However, when no parallel corpora are available, algorithms may resort to monolingual corpora and bilingual lexicon induction (Klementiev et al., 2012). The induction process requires a seed dictionary that GLAWI could provide for many language pairs. GLAWI's translations could also be used to complement existing multilingual resources such as PanDictionary (Mausam et al., 2009), a multilingual translation graphs which compiles numerous dictionaries. Translations may even help infer monolingual information. For example, they can be used to compute semantic relatedness: two words of a given language translating to the same words in different languages are likely to have close meanings (Sajous et al., 2013).

2.6. Morphological relations

GLAWI contains compounds, derivative and "related" words that correspond to Wiktionnaire's sections entitled *Composés*, *Dérivés* and *Apparentés étymologiques*. Examples of such morphological relations are presented for the noun *nom* 'name/noun' in Figure 7.

```
<subsection type="morpho">
  <item type="compound">nom commun</item>
  <item type="compound">nom collectif</item>
  <item type="compound">prête-nom</item>
  <item type="derivative">nommer</item>
  <item type="derivative">nommage</item>
  <item type="derivative">nomination</item>
  <item type="related">anonyme</item>
</subsection>
```

Figure 7: Morphological relations for *nom* 'name/noun'

The number of morphological sections per POS and the total number of morphological relations are given in Table 6. In addition to the morphological sections, information about derivational or compositional coinage of words may be found in the etymology sections (cf. section 2.3.). GLAWI's morphological relations may be used for research in computational morphology and to build morphological resources like Morphonette,[3] a paradigm-based morphological network (Hathout, 2011) and Démonette,[4] a French derivational morpho-semantic network (Hathout and Namer, 2014). They could also be leveraged in NLP applications. For example, Padó et al. (2013) use derivative words to overcome data sparseness in distributional analysis.

Morphological Relations			
POS	**# sections**	**Relations**	
nouns	16,948		
		compound	1,118
		derivative	50,506
		related	22,874
adjectives	4,939		
		compound	309
		derivative	9,481
		related	6,767
verbs	5,443		
		compound	109
		derivative	10,684
		related	5,170
adverbs	899		
		derivative	488
		related	1,284

Table 6: Morphological relations

2.7. Forms variation

In Wiktionnaire, alternative spellings may result in separate pages for the same word, such as *nénuphar* and *nénufar* 'water lily'. Other form variations result in redirection links. Though most of them only serve navigational purpose (e.g. to redirect to an existing page when ligatures or diacritics are omitted, when alternative single quotes are used, etc.). Some may be collected to build a lexicon of form variants (see Table 7). Moreover, common misspellings can be used by spell-checkers or for educational purposes. Alternative forms can also benefit text normalization in corpus processing and information retrieval. More deviant variations (oral transcriptions, texto language, etc.) can help analyze computer-mediated communications (Melero et al., 2012; Baldwin et al., 2013).

2.8. Phonemic transcriptions

94 % of GLAWI's entries contain one or several phonemic transcriptions. They may include diatopic variations. Figure 8a illustrates regional variants: *moins* 'minus, less'

[3] http://redac.univ-tlse2.fr/lexicons/morphonette_en.html

[4] http://redac.univ-tlse2.fr/lexicons/demonette_en.html

Form	Standard/other form	Translation/indication	Variation type
nénuphar	nénufar	water lily	alternative spelling suggested
maitriser	maîtriser	to master	by the 1990 reform
quinquenat	quinquennat	five year period	frequent mispelling
évidament	évidemment	obviously	
enkikiner	enquiquiner	to bother, to annoy	voluntary mispelling (texto/forum)
c'qui	ce qui	which	oral transcription
coeur	cœur	heart	ligature
& al.	et al.		symbol/litteral
VOIP	VoIP		case
écart type	écart-type	standard deviation	compound linking character
copier-coller	copier/coller	copy and paste	
climato-sceptique	climatosceptique	climateskeptics	
abreuvement	abreuvage	watering	concurrent suffixes
graticiel	gratuiciel	freeware	portmanteau formation
débit de boisson	débit de boissons	public house, pub	inflection within MWE
erratums	errata		French/Latin inflection
coulibiac	koulibiak	stuffed Russian baked dough	loan word/conventional
halal	hallal		transcriptions
mozzarelle	mozzarella	Italian mild cheese	
chai	tchaï	black tea	
clubbeur	clubber		
N'Djaména	Ndjamena		foreign proper name

Table 7: Examples of form variations

is pronounced /mwɛ̃/ in "standard" French (Paris) and /mwɛ̃s/ in Southern France (Marseille). An example of national variations is given in the Figure 8b, where two different transcriptions are given for *sorcière* 'witch' in France and Québec (Canada). Hathout et al. (2014b) have shown that the quality of Wiktionnaire's transcriptions and syllabation is comparable to those of existing phonological lexicons, the latter having a smaller coverage.

Quality pronunciation lexicons have a significant impact on text-to-speech systems. While unknown words are processed by machine-learned models, grapheme-to-phoneme conversion of common words use large-scale pronunciation lexicons (Rojc and Kačič, 2007). Phonemic transcriptions and syllabations are also widely used in psycholinguistics to set up experimental material for semantic priming, as in (Bracco et al., 2015).

```
<pronunciations>
  <pron area="Paris">mwɛ̃</pron>
  <pron area="Marseille">mwɛ̃s</pron>
</pronunciations>
```

(a) Transcriptions of *moins* 'minus, less'

```
<pronunciations>
  <pron area="France">sɔʁ.sjɛʁ</pron>
  <pron area="Québec">sɔʁ.sjaɛʁ</pron>
</pronunciations>
```

(b) Transcriptions of *sorcière* 'witch'

Figure 8: Examples of phonemic transcriptions with diatopic variations

3. Conclusion

The GLAWI machine-readable dictionary is a new type of lexicographical resource that eases the use of Wiktionary for both linguistic research and NLP. The standardization of the wikicode allows the user to easily extract a variety of information, such as neologisms, feminine equivalent of masculine nouns, etc. To date, it is the only free resource available for contemporary French that contains definitions. We plan to develop a user interface to query GLAWI by setting conditions on the individual fields that make up the entries in a way similar to GLÀFFOLI,[5] the online interface provided to manually query GLÀFF.

This work opens the way to the creation of similar resources for other languages, including those who do not yet have any freely available Machine-Readable Dictionary. Electronic dictionaries similar to GLAWI are under development for Italian and English. The many possible uses of this type of dictionaries will also improve the endowment of poorly or lesser-resourced languages in quality linguistic resources. At the time of writing, a morphosyntactic Serbian lexicon extracted from Wiktionary is currently being released.

4. Acknowledgements

Computations performed to produce GLAWI have been carried out using the OSIRIM platform, that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government and ERDF.

[5] http://redac.univ-tlse2.fr/glaffoli/

5. Bibliographical References

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How Noisy Social Media Text, How Different Social Media Sources? In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013*, pages 356–364, Nagoya, Japan.

Bracco, G., Calderone, B., and Celata, C. (2015). Phonotactic probabilities in Italian simplex and complex words: a fragment priming study. In *Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage: Representations and Processes in the Mental Lexicon*, pages 24–28, Pisa, Italy.

Cutler, M., Shih, Y., and Meng, W. (1997). Using the Structure of HTML Documents to Improve Retrieval. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, pages 241–252, Monterey, California.

Desalle, Y., Gaume, B., and Duvignau, K. (2009). SLAM. Automatic lexical solutions for metaphors. *TAL*, 50(1):145–175.

Desalle, Y., Gaume, B., Magistry, P., Duvignau, K., Cheung, H., and Hsieh, S.-K. (2014). Skillex: An Action Labelling Efficiency Score. The Case for French and Mandarin. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci2014)*, pages 409–414, Quebec City, Canada.

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

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, Ireland. Association for Computational Linguistics and Dublin City University.

Hathout, N., Sajous, F., and Calderone, B. (2014b). GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1007–1012, Reykjavik, Iceland.

Hathout, N. (2011). Morphonette: a paradigm-based morphological network. *Lingue e linguaggio*, 2011(2):243–262.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2012)*, pages 130–140, Avignon, France. Association for Computational Linguistics.

Mausam, Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., and Bilmes, J. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pages 262–270, Suntec, Singapore.

McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Melero, M., Costa-Jussà, M. R., Domingo, J., Marquina, M., and Quixal, M. (2012). Holaaa!! writin like u talk is kewl but kinda hard 4 NLP. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3794–3800, Istanbul, Turkey.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL 2007 Shared Task on dependency parsing (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic.

Padó, S., Šnajder, J., and Zeller, B. (2013). Derivational Smoothing for Syntactic Distributional Semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 731–735, Sofia, Bulgaria.

Rojc, M. and Kačič, Z. (2007). Time and Space-efficient Architecture for a Corpus-based Text-to-speech Synthesis System. *Speech Commununication*, 49(3):230–249.

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

Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Hrafn Loftsson, et al., editors, *Advances in Natural Language Processing*, volume 6233 of *LNCS*, pages 332–344. Springer Berlin / Heidelberg.

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

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, Netherlands.

Tiedemann, J. and Ljubešić, N. (2012). Efficient Discrimination Between Closely Related Languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

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.

Weale, T., Brew, C., and Fosler-Lussier, E. (2009). Using the Wiktionary Graph Structure for Synonym Detection. In *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–31, Suntec, Singapore.