From GLÀFF to PsychoGLÀFF: a large psycholinguistics-oriented French lexical resource

Basilio Calderone, Nabil Hathout, Franck Sajous
CLLE-ERSS (CNRS&Université de Toulouse-Le Mirail)
E-mail: {basilio.calderone, nabil.hathout, franck.sajous}@univ-tlse2.fr

Abstract

In this paper, we present two French lexical resources, GLÀFF and PsychoGLÀFF. The former, automatically extracted from the collaborative online dictionary Wiktionary, is a large-scale versatile lexicon exploitable in Natural Language Processing applications and linguistic studies. The latter, based on GLÀFF, is a lexicon specifically designed for psycholinguistic research.

GLÀFF, counting more than 1.4 million entries, features an unprecedented size. It reports lemmas, main syntactic categories, inflectional features and phonemic transcriptions. PsychoGLÀFF contains additional information related to formal aspects of the lexicon and its distribution. It contains about 340,000 entries (120,000 lemmas) that are corpora-attested. We explain how the resources have been created and compare them to other known resources in terms of coverage and quality. Regarding PsychoGLÀFF, the comparison shows that it has an exceptionally large repertoire while having a comparable quality.

Keywords: French lexicon; lexical resource for psycholinguistic studies; Wiktionary

1 Introduction

Lexical resources play an important role in psycholinguistics by providing researchers with a set of experimentally relevant corpus information concerning words and, to a lesser extent, their sub-lexical components. In particular, psycholinguists working on lexical access need to manipulate a set of formal properties of words, such as syllabification, phonemic transcription, lemmas, inflected forms or orthographic/phonological neighborhood (i.e., the number of words differing from the target word by only one character/phoneme). Word frequency is possibly the most crucial information to be accounted for in psycholinguistic studies, and it is generally provided for either wordforms, lemmas, or both. The most well-known resource for English, German and Dutch is probably CELEX (Baayen et al. 1995). Many other languages, including French, lack a similar resource.

Some freely available French morphological lexicons, such as Lefff (Clément et al. 2004) and Morphalou (Romary et al. 2004), contain inflected forms, lemmas and morphosyntactic tags. These resources, designed in the first place for natural language processing (NLP) or lexicography do not include, however, phonemic transcriptions that are necessary to set up psycholinguistics experiments, for extensive morphology and in the design of tools such as phonetizers. One noticeable exception is Lexique (New 2006), a free lexicon quite popular in psycholinguistics. This lexicon includes phonemic transcriptions, word frequencies and various features relevant to this field. However it has a limited coverage, especially in terms of inflected forms. All other resources that have both exploitable coverage and phonemic transcriptions, such as BDLex (Pérennou and de Calmès 1987), ILPho (Boula De Mareuil et al. 2000) or GlobalPhone (Schultz et al. 2013) are not free. Besides their cost, derivative works cannot be redistributed, which constitutes an impediment for collaborative research. As of today, no French lexicon meets all following requirements: free license, wide coverage, phonemic transcriptions and word frequencies.

In this article, we present a psycholinguistics-oriented resource based on *Wiktionnaire*,[1] the French edition of Wiktionary. In a previous work (Sajous et al. 2013a), we automatically extracted GLÀFF, *"a Large Versatile French Lexicon"*. This large-scale resource contains, for each entry, inflectional and phonemic information. PsychoGLÀFF is a new step leveraging Wiktionnaire's content.[2] Grounded on GLÀFF, PsychoGLÀFF is a lexicon that contains additional features specifically designed for meeting psycholinguistic needs.

The paper is organized as follows: Section 2 gives an overview of Wiktionnaire and its features relevant to lexical resources building. GLÀFF is then described in Section 3. Finally, we present in Section 4 PsychoGLÀFF, a lexicon designed for psycholinguistics use and compare it to Lexique in terms of coverage and word frequency. Conclusions and future directions of work are discussed in Section 5.

2 Wiktionnaire as a source of lexical knowledge

Wiktionary is a free multilingual dictionary available online. As its mother project Wikipedia, Wiktionary is based on the wiki paradigm: every internet user may contribute by adding content or modifying existing one. Launched in 2003, the Wiktionary project boasts, ten years later, more than two million entries for its French language edition, the Wiktionnaire. The impressive size of its headword list has to be tempered: inflected forms, discussion pages and, more surprisingly, *"pages describing in French words from other languages"* are counted as regular entries. However, once these latter entries excluded, Wiktionnaire still accounts for 1.4 million entries (186,000 lemmas).

While Wikipedia has been extensively used in various disciplines, its lexicographic counterpart seems to have received

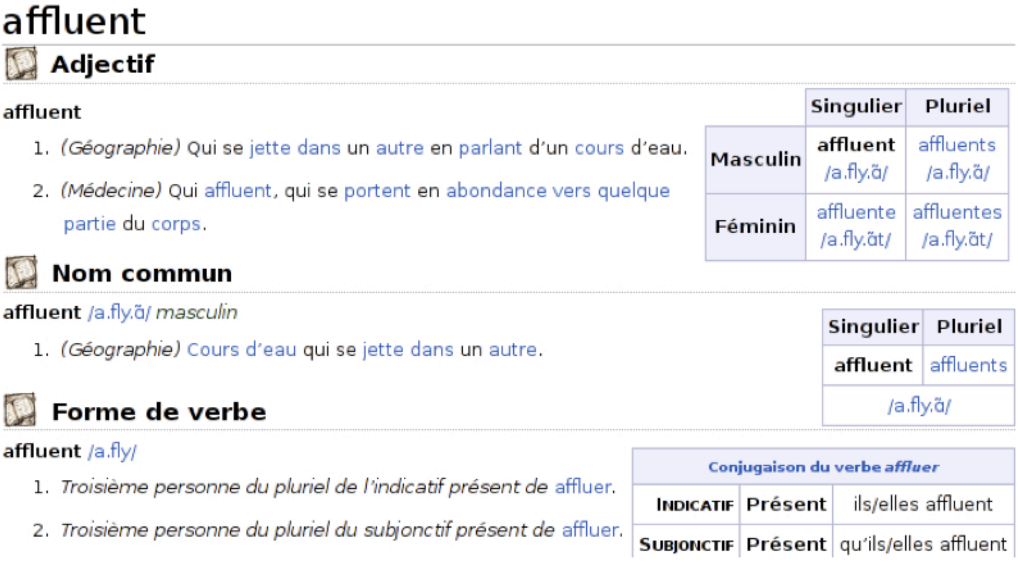
[1] http://fr.wiktionary.org

[2] GLÀFF and PsychoGLÀFF are freely available from the REDAC website: http://redac.univ-tlse2.fr/lexicons/

less attention from the scientific community. Wiktionary was first used in NLP by Zesch et al. (2008) to compute semantic relatedness. Its potential as an electronic lexicon was studied for the first time by Navarro et al. (2009) for French and English synonymy mining. Along the same line of research, Anton Perez et al. (2011) realized the integration of the Portuguese edition of Wiktionary in the ontology Onto.PT (Gonçalo Oliveira and Gomes 2010). Serasset (2012) designed Dbnary, an open-source resource containing "easily extractable entries". The French subpart of this resource contains 260,467 entries. Works led by Meyer and Gurevych (2012) and Gurevych et al. (2012) resulted in German and English ontologies based on Wiktionary. Sajous et al. (2010) has made available a structured XML version of this lexicon for French and English, called WiktionaryX.[3]

Although Wiktionnaire presents interesting features (unprecedented coverage, definitions, phonemic transcriptions, semantic relations, translations, free license),[4] the information it contains is difficult to extract. This probably explains the relatively small number of works using it. Wiktionnaire, as other Wiktionary's language editions, is released as an "XML dump", where XML only marks the macrostructure. The microstructure is encoded in a format called *wikicode*, inherent in the content management system *MediaWiki*. This format has no formally defined syntax, evolves over time, and is not stable from one language edition to another. This underspecified syntax makes therefore the automatic information extraction from the collaborative dictionary uneasy: multiple deviations from a "prototypical article" should be expected, as well as missing information, redundancy and inconsistency.

Figure 1 shows the entry *affluent* (adjective and noun 'affluent', and two inflection forms of the verb *affluer* 'to flow into/to pour in') as it is visible in Wiktionnaire. The corresponding wikicode of this article is shown in Figure 2. Inflected forms may appear in the article related to their lemma (as it is the case in Figure 1). They may also have a dedicated page (cf. Figures 3 and 4).



affluent

Adjectif

affluent

1. (Géographie) Qui se jette dans un autre en parlant d'un cours d'eau.
2. (Médecine) Qui affluent, qui se portent en abondance vers quelque partie du corps.

	Singulier	Pluriel
Masculin	affluent /a.fly.ɑ̃/	affluents /a.fly.ɑ̃/
Féminin	affluente /a.fly.ɑ̃t/	affluentes /a.fly.ɑ̃t/

Nom commun

affluent /a.fly.ɑ̃/ masculin

1. (Géographie) Cours d'eau qui se jette dans un autre.

Singulier	Pluriel
affluent	affluents
/a.fly.ɑ̃/	

Forme de verbe

affluent /a.fly/

1. Troisième personne du pluriel de l'indicatif présent de affluer.
2. Troisième personne du pluriel du subjonctif présent de affluer.

Conjugaison du verbe *affluer*		
Indicatif	Présent	ils/elles affluent
Subjonctif	Présent	qu'ils/elles affluent

Figure 1: Article of the word *'affluent'* in Wiktionnaire

The table of the adjective inflected forms (top-right in Figure 1) is not explicitly present in the wikicode, but is generated by the template `{{fr-accord-cons|a.fly.ɑ̃|t}}` (cf. Figure 2). There are hundreds of similar patterns in the wikicode. An example of the non-systematic wikicode's format and resulting article's layout can be seen in Figure 3: unlike the template `{{f}}` that defines the feminine gender of the form, there is no template specifying the grammatical number. The number can only be extracted by parsing the definition *"Féminin singulier"*. The heterogeneity of the wikicode also concerns the phonemic transcriptions: they occur sometimes in the *Ligne de forme* (the line following the part of speech heading), as in Figure 3 for *'affluente'*, and sometimes, on the contrary, they are specified in a separate *"Prononciation"* section as in Figure 4 for *'affluentes'*.

To build GLÀFF and PsychoGLÀFF, we automatically extracted the inflected forms and lemmas in their dedicated pages, and detected the inflection templates. We also identified the phonemic transcriptions wherever they occur. We finally parsed the conjugation tables (cf. Figure 5). We thus collected as much (possibly redundant) information as possible and applied some heuristics to automatically detect major inconsistencies.

[3] WiktionaryX, is an XML version of the English and French editions of Wiktionary, freely available at http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html

[4] For a more comprehensive description of the Wiktionnaire, see (Navarro et al. 2009) and (Sajous et al. 2010; 2013b).

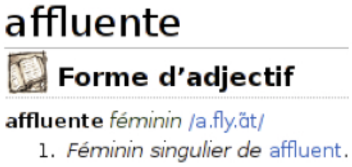
```
{{-adj-|fr}}
{{fr-accord-cons|a.fly.ɑ̃|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]]
d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ɑ̃}}

{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui|}}
'''affluent''' {{pron|a.fly|fr}}
# ''3ème pers. du pluriel de l'indicatif présent de'' [[affluer]].
# ''3ème pers. du pluriel du subjonctif présent de'' [[affluer]].

{{-pron-}}
{| class="wikitable"
| Adjectif et nom commun
* {{pron-rég|France|ɛ̃ .n‿a.fly.ɑ̃|titre=un affluent}}
|-
| Forme du verber affluer
* {{pron-rég|France (Île-de-France)|a.fly}}
```

Figure 2: Wikicode of the article *'affluent'*



affluente

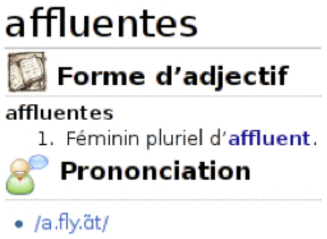
Forme d'adjectif

affluente *féminin* /a.fly.ɑ̃t/

1. *Féminin singulier de* affluent.

```
{{-flex-adj-|fr}}
'''affluente''' {{f}} {{pron|a.fly.ɑ̃t|lang=fr}}
#''Féminin singulier de'' [[affluent#fr-adj|affluent]].
```

Figure 3: Article and wikicode of *'affluente'*



affluentes

Forme d'adjectif

affluentes

1. Féminin pluriel d'**affluent**.

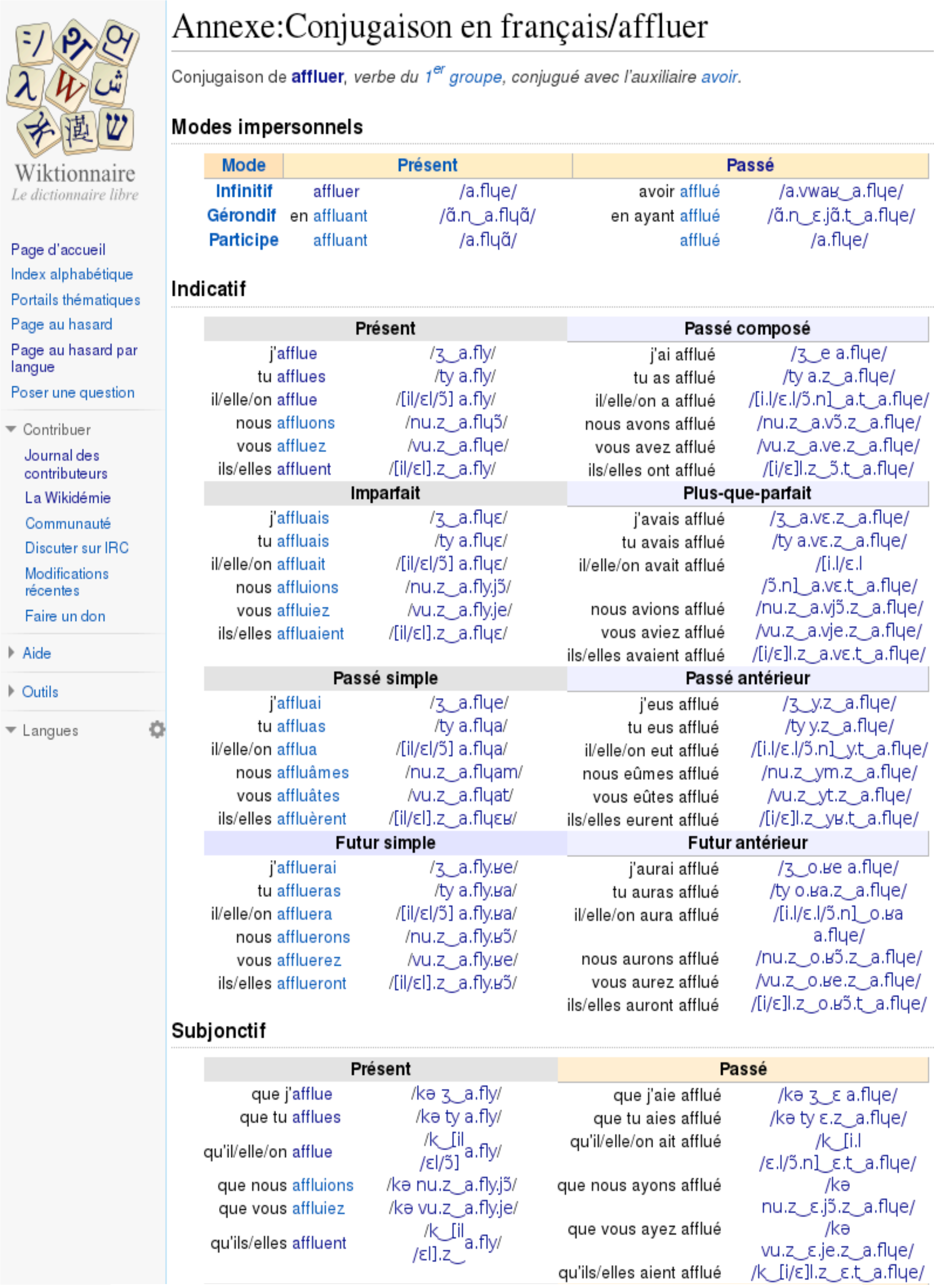
Prononciation

- /a.fly.ɑ̃t/

```
{{-flex-adj-|fr}}
'''affluentes'''
# Féminin pluriel d''''[[affluent]]'''.

{{-pron-}}
* {{pron|a.fly.ɑ̃t}}
```

Figure 4: Article and wikicode of *'affluentes'*



Annexe:Conjugaison en français/affluer

Conjugaison de **affluer**, *verbe du 1er groupe, conjugué avec l'auxiliaire avoir.*

Modes impersonnels

Mode	Présent		Passé	
Infinitif	affluer	/a.flɥe/	avoir afflué	/a.vwaʁ‿a.flɥe/
Gérondif	en affluant	/ɑ̃.n‿a.flɥɑ̃/	en ayant afflué	/ɑ̃.n‿ɛ.jɑ̃.t‿a.flɥe/
Participe	affluant	/a.flɥɑ̃/	afflué	/a.flɥe/

Indicatif

Présent		Passé composé	
j'afflue	/ʒ‿a.fly/	j'ai afflué	/ʒ‿e a.flɥe/
tu afflues	/ty a.fly/	tu as afflué	/ty a.z‿a.flɥe/
il/elle/on afflue	/[il/ɛl/ɔ̃] a.fly/	il/elle/on a afflué	/[i.l/ɛ.l/ɔ̃.n]‿a.t‿a.flɥe/
nous affluons	/nu.z‿a.flɥɔ̃/	nous avons afflué	/nu.z‿a.vɔ̃.z‿a.flɥe/
vous affluez	/vu.z‿a.flɥe/	vous avez afflué	/vu.z‿a.ve.z‿a.flɥe/
ils/elles affluent	/[il/ɛl].z‿a.fly/	ils/elles ont afflué	/[i/ɛ]l.z‿ɔ̃.t‿a.flɥe/
Imparfait		**Plus-que-parfait**	
j'affluais	/ʒ‿a.flɥɛ/	j'avais afflué	/ʒ‿a.vɛ.z‿a.flɥe/
tu affluais	/ty a.flɥɛ/	tu avais afflué	/ty a.vɛ.z‿a.flɥe/
il/elle/on affluait	/[il/ɛl/ɔ̃] a.flɥɛ/	il/elle/on avait afflué	/[i.l/ɛ.l/ɔ̃.n]‿a.vɛ.t‿a.flɥe/
nous affluions	/nu.z‿a.fly.jɔ̃/	nous avions afflué	/nu.z‿a.vjɔ̃.z‿a.flɥe/
vous affluiez	/vu.z‿a.fly.je/	vous aviez afflué	/vu.z‿a.vje.z‿a.flɥe/
ils/elles affluaient	/[il/ɛl].z‿a.flɥɛ/	ils/elles avaient afflué	/[i/ɛ]l.z‿a.vɛ.t‿a.flɥe/
Passé simple		**Passé antérieur**	
j'affluai	/ʒ‿a.flɥe/	j'eus afflué	/ʒ‿y.z‿a.flɥe/
tu affluas	/ty a.flɥa/	tu eus afflué	/ty y.z‿a.flɥe/
il/elle/on afflua	/[il/ɛl/ɔ̃] a.flɥa/	il/elle/on eut afflué	/[i.l/ɛ.l/ɔ̃.n]‿y.t‿a.flɥe/
nous affluâmes	/nu.z‿a.flɥam/	nous eûmes afflué	/nu.z‿ym.z‿a.flɥe/
vous affluâtes	/vu.z‿a.flɥat/	vous eûtes afflué	/vu.z‿yt.z‿a.flɥe/
ils/elles affluèrent	/[il/ɛl].z‿a.flɥɛʁ/	ils/elles eurent afflué	/[i/ɛ]l.z‿yʁ.t‿a.flɥe/
Futur simple		**Futur antérieur**	
j'affluerai	/ʒ‿a.fly.ʁe/	j'aurai afflué	/ʒ‿o.ʁe a.flɥe/
tu afflueras	/ty a.fly.ʁa/	tu auras afflué	/ty o.ʁa.z‿a.flɥe/
il/elle/on affluera	/[il/ɛl/ɔ̃] a.fly.ʁa/	il/elle/on aura afflué	/[i.l/ɛ.l/ɔ̃.n]‿o.ʁa a.flɥe/
nous affluerons	/nu.z‿a.fly.ʁɔ̃/	nous aurons afflué	/nu.z‿o.ʁɔ̃.z‿a.flɥe/
vous affluerez	/vu.z‿a.fly.ʁe/	vous aurez afflué	/vu.z‿o.ʁe.z‿a.flɥe/
ils/elles afflueront	/[il/ɛl].z‿a.fly.ʁɔ̃/	ils/elles auront afflué	/[i/ɛ]l.z‿o.ʁɔ̃.t‿a.flɥe/

Subjonctif

Présent		Passé	
que j'afflue	/kə ʒ‿a.fly/	que j'aie afflué	/kə ʒ‿ɛ a.flɥe/
que tu afflues	/kə ty a.fly/	que tu aies afflué	/kə ty ɛ.z‿a.flɥe/
qu'il/elle/on afflue	/k‿[il/ɛl/ɔ̃] a.fly/	qu'il/elle/on ait afflué	/k‿[i.l/ɛ.l/ɔ̃.n]‿ɛ.t‿a.flɥe/
que nous affluions	/kə nu.z‿a.fly.jɔ̃/	que nous ayons afflué	/kə nu.z‿ɛ.jɔ̃.z‿a.flɥe/
que vous affluiez	/kə vu.z‿a.fly.je/	que vous ayez afflué	/kə vu.z‿ɛ.je.z‿a.flɥe/
qu'ils/elles affluent	/k‿[il/ɛl].z‿a.fly/	qu'ils/elles aient afflué	/k‿[i/ɛ]l.z‿ɛ.t‿a.flɥe/

Figure 5. Conjugation table of the verb *affluer* (extract)

3 GLÀFF

In this section, we summarize some relevant characteristics of GLÀFF, first introduced in (Sajous et al. 2013a), from which PsychoGLÀFF is derived. The latest version of GLÀFF includes nouns, verbs, adjectives, adverbs, and function words. As can be seen in Figure 6, GLÀFF specifies for each entry:

- the wordform;
- the lemma;
- the part of speech and morphosyntactic features in GRACE format (Rajman et al. 1997);
- the phonological transcription(s) (when specified in Wiktionnaire) in IPA and in SAMPA with syllable boundaries.

```
affluent|Ncms|affluent|a.fly.ɑ̃|a.fly.A~
affluent|Afpms|affluent|a.fly.ɑ̃|a.fly.A~
affluents|Afpmp|affluent|a.fly.ɑ̃|a.fly.A~
affluents|Ncmp|affluent|a.fly.ɑ̃|a.fly.A~
affluent|Vmip3p-|affluer|a.fly|a.fly
affluent|Vmsp3p-|affluer|a.fly|a.fly
```

Figure 6: Extract of GLÀFF

3.1 Coverage

GLÀFF differs from the lexicons currently used in NLP and psycholinguistics by its exceptional size. Table 1 shows the number of inflected forms and lemmas for simple words (only letters) and non-simple words (containing spaces, dashes or digits) in five different French lexicons. GLÀFF contains from 3 to 4 times more tokens and from 3 to 9 times more inflected forms than the other lexicons.

	Categorized inflected forms			Categorized lemmas		
	Simple	Non simple	Total	Simple	Non simple	Total
Lexique	147,912	4,696	152,608	46,649	3,770	50,419
BDLex	431,992	4,360	436,352	47,314	1,792	49,106
Lefff	466,668	3,829	470,497	54,214	2,303	56,517
Morphalou	524,179	49	524,228	65,170	7	65,177
GLÀFF	1,401,578	24,270	1,425,848	172,616	13,466	186,082

Table 1: Size of five French lexicons (counting only nouns, verbs, adjectives and adverbs).

Table 2 reports the intersection of GLÀFF with the other lexicons. We observe that the magnitude of the intersection depends on the size of the lexicons: the bigger a lexicon, the larger its intersection with the other ones. Three groupings emerge: Lexique has the smallest coverage, only containing 9% of GLÀFF and 22% to 26% of the entries of the other lexicons. BDLex, Lefff and Morphalou cover 76% to 80% of Lexique and about 30% of GLÀFF. Finally GLÀFF is clearly on top with coverage of 85% to 93%. In total, its coverage is 5% to 65% higher than the other lexicons.

	Lexique	BDLex	Lefff	Morphalou	GLÀFF
Lexique	-	26.03	25.20	22.46	8.95
BDLex	76.02	-	79.87	70.40	28.75
Lefff	79.50	86.28	-	72.32	30.04
Morphalou	79.58	85.43	81.24	-	32.03
GLÀFF	84.83	93.26	90.23	85.66	-

Table 2: Intersection of five French lexicons (% of the categorized inflected forms).

Size is a crucial aspect of the lexicons used for research in derivational and inflectional morphology or, more generally, in the development of NLP tools such as morphosyntactic taggers and parsers. In order to asses that GLÀFF's largest size is actually useful, we compared the five lexicons with the vocabulary of four corpora of various types. Frantext 20^e is constituted by 515 novels of 20th century French literature containing 30 million words. LM10 is a 200 million word corpus made up of the archives of the newspaper Le Monde from 1991 to 2000. The third corpus, containing 260 million words, consists of the articles from the French Wikipedia. Finally, FrWaC (Baroni et al. 2009) is a 1.6 billion words corpus of French web pages. Table 3 shows the coverage of the five lexicons with respect to the four corpora.

Threshold: frequency ≥		**1**	**2**	**5**	**10**	**100**	**1000**
Frantext	#forms	145,437	95,189	61,813	43,919	10,767	1,376
	Lexique	66.76	84.35	94.00	**96. 91**	**99.15**	**99.27**
	BDLex	70.86	84.69	92.47	95.74	99.12	99.20
	Lefff	71.89	85.63	93.21	96.21	99.08	98.90
	Morphalou	73.93	86.66	93.29	96.00	98.48	97.09
	GLÀFF	**76.92**	**88.57**	**94.54**	96.72	98.77	98.76
LM10	#forms	300,606	172,036	106,470	77,936	29,388	83.21
	Lexique	29.59	47.28	65.23	76.31	93.81	98.58
	BDLex	37.77	55.79	71.76	80.93	95.53	98.69
	Lefff	39.64	58.22	74.33	83.20	95.99	**98.90**
	Morphalou	39.06	56.82	71.92	80.32	93.27	97.48
	GLÀFF	**45.24**	**63.83**	**78.63**	**86.23**	**96.46**	98.68
Wikipedia	#forms	953,920	435,031	216,210	136,531	35,621	7,956
	Lexique	9.13	18.27	31.52	43.03	78.58	95.72
	BDLex	12.29	22.89	36.80	48.04	79.39	95.33
	Lefff	12.88	23.94	38.26	49.65	80.57	95.71
	Morphalou	13.05	23.96	37.87	48.87	78.74	94.16
	GLÀFF	**16.42**	**29.00**	**44.13**	**55.45**	**83.21**	**96.10**
FrWaC	#forms	1,624,620	846,019	410,382	255,718	74,745	22,100
	Lexique	5.83	10.85	20.84	30.81	66.00	89.47
	BDLex	9.36	15.85	27.28	37.48	69.61	90.03
	Lefff	9.85	16.67	28.57	39.16	71.61	91.16
	Morphalou	10.09	16.89	28.53	38.68	69.36	88.51
	GLÀFF	13.13	21.13	34.29	45.35	76.39	92.76

Table 3: Lexicon/corpus coverage (% of non-categorized inflected forms).

The vocabulary is restricted to the forms that occur at least once, 2, 5, 10, 100 and 1000 times. The ranking of the corpora by coverage is the same for the five lexicons. Although their size affects the order, their nature is also crucial. For example, FrWaC being a collection of web pages, it contains a large number of "noisy" forms (foreign words, missing or extra spaces, missing diacritics, random spelling, etc.). Again, we see the division of lexicons into three groups. BDLex, Lefff and Morphalou have a quite close coverage. Except for Frantext 20^{e}, Lexique has the smallest coverage. GLÀFF has the largest coverage for all corpora, except for LM10 at the 1000 threshold where it is surpassed by Lefff by 0.2%. The best coverage of Lexique for the Frantext 20^{e} corpus, above the 10 threshold, while it has the weakest coverage in all other cases, is explained by the design of its vocabulary, extracted from this corpus. For the other corpora and up to the 100 threshold, the size of GLÀFF explains its larger coverage with respect to the other lexicons (at the threshold 1, 14% to 53% larger for LM10 and 30% to 120% larger for FrWaC; at the threshold 10, 4% to 16% for LM10 and 15% to 47% for FrWaC). NLP tools that integrate GLÀFF should therefore offer an improved performance in the treatment of these corpora. In a qualitative study described in (Sajous et al. 2014), we observed that GLÀFF specific entries contains not only rare neologisms, but also very common words such as *attractivité* 'attractivity', *brevetabilité* 'patentability', *diabolisation* 'demonization', *employabilité* 'employability', *homophobie* 'homophobia', *hébergeur* 'host', fatwa, *institutionnellement* 'institutionally', *anticorruption* 'anti-corruption', etc. missing from the other lexicons.

3.2 Phonemic transcriptions

GLÀFF provides a phonemic transcription for about 90% of the entries. We evaluated the consistency of these transcriptions with respect to those of BDLex and Lexique (after conversion into IPA encoding).

Tables 4a to 4c report the top ten variations between pairs from the three lexicons. We only considered one phoneme differences, ignoring syllabification. The differences in transcriptions between GLÀFF and the other two lexicons are comparable to the differences observed between BDLex and Lexique. In particular, these differences are mostly due to the distinctions between the mid vowels, i.e. the front-mid vowels: [e] (close-mid) vs. [ɛ] (open-mid) and the back-mid vowels: [o] (close-mid) vs. [ɔ] (open-mid). This alternation is a well-known aspect of French phonology resulting from diatopic variations (North vs. South), as described in (Detey et al. 2010). Such expected oppositions account for about 91% of the divergences between BDLex and Lexique. Table 5 reports the percentage of identical phonological transcriptions shared by the lexicons and the percentage of the 'comparable' phonological transcriptions, i.e. disregarding the distinction between close-mid and open-mid vowels. GLÀFF and Lexique give identical transcriptions for 79.5% of entries whereas the percentage between GLÀFF and BDLex is lower, at 61.7%. Table 5 also reports the results of the comparison of syllabification in the three lexicons (performed on the basis of identical transcriptions only). This comparison shows that the three lexicons are quite similar with respect to syllabification (98%).

Comparing GLÀFF with the major resources that contain the same type of information clearly shows that the overall quality of the lexicon is quite satisfactory and is in all respect comparable to those of these resources.

Oper.	Phonemes	%	∑ %
r	ɛ/e	48.18	48.18
r	ɔ/o	32.17	80.36
r	o/ɔ	11.02	91.37
r	y/ɥ	1.83	93.21
r	ə/ø	1.44	94.64
r	ə/œ	1.39	96.03
r	u/w	0.84	96.87
r	b/p	0.73	97.61
r	s/z	0.51	98.12
d	j	0.25	98.37

(a) BDLex/Lexique

Oper.	Phonemes	%	∑ %
r	ɔ/o	60.03	60.03
i	ə	14.18	74.21
r	e/ɛ	6.90	81.11
r	ɛ/e	4.98	86.09
r	ɑ/a	4.92	91.01
r	s/z	1.25	92.26
r	ə/ø	0.91	93.17
r	œ/ø	0.47	93.64
i	i	0.42	94.06
r	o/ɔ	0.38	94.44

(b) GLÀFF/Lexique

Oper.	Phonemes	%	∑ %
r	e/ɛ	66.46	66.46
r	ɔ/o	10.58	77.05
i	ə	5.90	82.96
r	o/ɔ	4.36	87.32
r	ɑ/a	3.84	91.17
r	ɥ/y	1.61	92.78
r	œ/ə	1.09	93.88
r	ø/ə	0.86	94.74
i	i	0.84	95.58
r	w/u	0.79	96.38

(c) GLÀFF/BDLex

Table 4: The 10 most frequent differences in phonemic transcriptions
(Operations: r = replacement, i = insertion, d = deletion)

			Phonemic transcriptions		Syllabification
Lexicons		Intersection	Identical	Comparable	Identical
BDLex	Lexique	112,439	58.31	96.88	98.92
GLÀFF	Lexique	123,630	79.50	97.81	98.48
GLÀFF	BDLex	396,114	61.72	96.88	98.30

Table 5: Inter-lexicon agreement: phonemic transcriptions and syllabification

4 From GLÀFF to PsychoGLÀFF

4.1 Overview

Our goal in creating PsychoGLÀFF is to provide psycholinguists with a set of features related to the formal aspects of the lexicon entries. For this purpose, we selected from GLÀFF only forms having non-zero frequency in at least one of the corpora mentioned in section 3.1. This means that PsychoGLÀFF only contains lexical entries attested in the corpora, amounting to about 340,000 forms for 120,000 lemmas.

In addition to GLÀFF's features, PsychoGLÀFF includes the following information for each entry:

- the absolute and relative frequencies of the wordform and of the lemma in the aforementioned French corpora (Frantext 20^e, LM10 and FrWac);
- the length of the wordform (number of characters);
- the length of the phonological transcription(s) (number of phonemes);
- the syllabification and the CV structure of the wordform;
- the number of syllables;
- the geometric mean of the conditional character probabilities of bigrams, which calculates the probability of the bigram occurring given the preceding bigram;
- the geometric mean of the conditional character probabilities of trigrams, which calculates the probability of the trigram occurring given the preceding trigram;
- the geometric mean of the conditional character probabilities of 4-grams, which calculates the probability of the 4-gram occurring given the preceding 4-gram;
- the geometric mean of the conditional phoneme probabilities respectively calculated for bigrams, trigrams and 4-grams.
- the size of the orthographic neighbourhood, i.e. the number of words in the lexicon differing by one character (via deletion, insertion, or substitution);
- the size of the phonological neighborhood, i.e. the number of words differing from the phonological transcription by one phoneme (via deletion, insertion, or substitution);
- the size of the ratio between the number of consonants and syllables composing the phonological form. This score is meant to provide an estimate of the 'syllabic complexity' of the form.

The *n*-gram conditional probability represents a measure of phonotactic occurrence, defining the likelihood of occurrence of *n*-grams in French. This kind of measure is expected to be particularly helpful for the design of experimental stimuli in lexical access experiments (Storkel and Hoover 2011).

4.2 Comparison with Lexique

We compare hereafter, in terms of coverage and word frequencies, PsychoGLÀFF and Lexique, the most frequently used French lexicon in psycholinguistics.

Being directly extracted from GLÀFF, PsychoGLAFF stands out with respect to the lexicons currently used in psycholinguistics mostly for its size, as it counts 337,572 entries. Table 6 shows the number of inflected forms and lemmas of Lexique and PsychoGLAFF. The relative coverage of these lexicons is reported in Table 7.

	Categorized inflected forms	Categorized lemmas
Lexique	153,934	50,419
PsychoGLÀFF	337,572	121,021

Table 6: Size of Lexicons
(restricted to nouns, verbs, adjectives and adverbs)

	Lexique	PsychoGLÀFF
Lexique		36.1 %
PsychoGLÀFF	78.9 %	

Table 7: Lexicons relative coverage
(% of categorized inflected forms)

We observe that PsychoGLÀFF is more than twice larger than Lexique (for both inflected forms and lemmas) and has a total coverage of about 79% with respect to Lexique (which covers only 36.1% of the inflected forms of PsychoGLÀFF).

PsychoGLÀFF reports the absolute and relative frequencies for its wordforms and lemmas. Frequencies are calculated on the basis of three stylistically different corpora of written French: the abovementioned Frantext20^{e}, LM10 and FrWaC (literature, newspaper and web corpora). Lexique reports word frequency estimates too. It exploits two smaller corpora: a) a written corpus made up of 218 books from Frantext 20^{e}; b) a corpus of French subtitles for 9,474 movies and television series, assumed to be more representative of spoken French.

While GLÀFF and PsychoGLÀFF frequencies are exclusively based on written French, Lexique mixes together spoken-like and written resources for the calculation of wordform and lemma frequencies. Although the corpora used by the two lexicons have very different sizes, we attempted a comparison of the PsychoGLÀFF's frequencies with respect to the frequencies reported in Lexique, looking only at the intersection of the two lexicons.

Table 8 reports the correlation between the normalized frequencies of wordforms in PsychoGLÀFF (separately for Frantext, LM10 and FrWac) and Lexique (separately for books and movie subtitles). The data were normalized by one million words. It is not surprising that the correlation between Frantext's frequencies and Lexique's book frequencies is quite high (Pearson's coefficient ρ = .81), the latter being a sub-corpus of Frantext 20^{e}. Although PsychoGLÀFF frequencies are based exclusively on written corpora, we found a statistically significant correlation ρ = .68 between Lexique's subtitle frequencies and the Frantext frequencies (the value slightly decreases for the subtitles/FrWac correlation). This seems to indicate that the lexical coverage of PsychoGLÀFF, though based on written sources, is comparable to a relevant extent to the coverage of corpora specifically devoted to spoken French.

		Lexique	
		Subtitles	Books
PsychoGLÀFF	Frantext	.68	.81
	LM10	.62	.59
	FrWac	.67	.62

Table 8: PsychoGLÀFF/Lexique correlations with respect their normalized frequency values

An additional property of PsychoGLÀFF worth noting to is the presence of infrequent lexical items. This feature clearly derives from the nature of Wiktionnaire: being an online dictionary, it has not to conform to the same size constraints as printed ones. Bootstrapped by importation of articles from public domain dictionaries, it contains dated entries. Finally, being crowdsourced, it is regularly updated and contains general-domain neologisms, as well as subculture vocabulary and technical terms (Sajous et al. 2014). As a consequence, PsychoGLÀFF contains a large number of specific entries.

The bigger the corpus, the more low-frequency lexical items are likely to be included (while the size of the corpus is not likely to have a strong impact on the number of those words that are frequent or very frequent in a language, representing to a certain extent the 'essential lexicon' of that language). The Figure 7 illustrates this point by showing the distribution of different frequency intervals for both Lexique's and PsychoGLÀFF's sub-corpora. A normalized frequency range of 10.01 or more corresponds to very high words frequency and is situated at the right edge of the graph. A frequency range of 0.01-0.1 corresponds to very low words frequency and is situated at the left edge of the figure.

Six intermediate ranges capture the frequency differences of the entire lexicon. The figure shows that the distribution of the frequency intervals is approximately the same for Lexique and PsychoGLÀFF, with the significant exception of the least frequent word class (< 0.2), for which the number of lexical items in PsychoGLÀFF is almost twice as large as that of Lexique. At the same time, PsychoGLÀFF contains many words of ordinary usage that are absent from Lexique, such as *acceptabilité* 'acceptability', *centralité* 'centrality', *Saturne* 'Saturn', etc. In this sense PsychoGLÀFF offers a much larger lexical repertoire not only in terms of tokens, but also in terms of types, which represents a particularly interesting feature for psycholinguistic studies and corpus investigations.

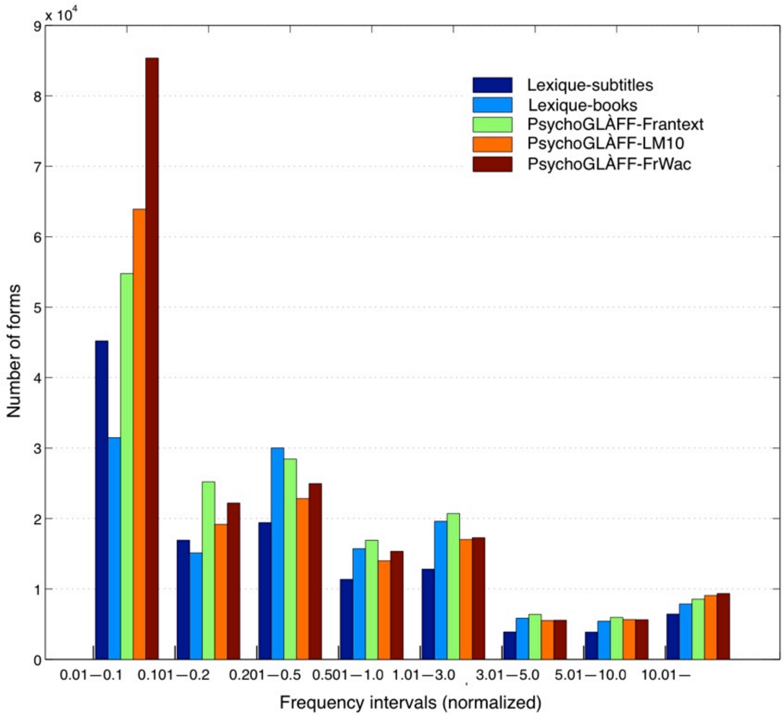


Figure 7: Distribution of forms with respect to their corpus frequency.

5 Conclusions and future directions

This paper presents a first version of PsychoGLÀFF, a large lexicon designed for psycholinguistic experimentation. PsychoGLÀFF was built on top of the inflectional and phonemic lexicon GLÀFF, itself acquired from Wiktionnaire, the French edition of the collaborative dictionary Wiktionary. In particular, PsychoGLÀFF contains the subset of GLÀFF's corpora-attested entries. This resource complements the inflectional and phonological information present in GLÀFF with features needed for experimental material calibration including frequency, lexical neighborhood, syllabic complexity and phonotactic likelihood.

Like GLÀFF, PsychoGLÀFF is characterized by an exceptional coverage, much higher than those of comparable resources as Lexique on one hand and Morphalou and Lefff on the other. We also show that the "primary" information (parts of speech, phonemic transcriptions, frequency) of PsychoGLÀFF and GLÀFF has a satisfactory quality. PsychoGLÀFF is a free resource distributed under a copylefted license and is available to all psycholinguistic researchers working on French. We hope that it will soon be adopted by this community whose feedback will allow us to improve the resource and appropriately respond to its needs.

In the near future, we plan to improve PsychoGLÀFF on several aspects. One of them will concern the description completeness and consistency of the lexicon. An online interface, comparable to GLÀFFOLI (the GLÀFF OnLine Interface) will also be developed, that will enable users to query the lexicon and develop experimental material interactively.

6 References

Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, pp. 703–717, Lisboa, Portugal.

Baayen, R. H., Piepenbrock, R. and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia: Linguistic Data Consortium.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3): 209-226.

Boula De Mareuil, P., Yvon, F., D'Alessandro, C., Aubergé, V., Vaissière, J., and Amelot, A. (2000). A French Phonetic Lexicon with variants for Speech and Language Processing. In *Proceedings of the Second International*

Conference on Language Resources and Evaluation (LREC 2000), pp. 273–276, Athens, Greece.

Clément, L., Lang, B. and Sagot. B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004),* pp. 1841–1844, Lisboa, Portugal.

Detey, S., Durand, J., Laks, B., and Lyche, C. (2010). *Les variétés du français parlé dans l'espace francophone*. Paris: Ophrys.

Gonçalo Oliveira, H. and Gomes, P. (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium*, pp. 199–211. IOS Press.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer C. M. and Wirth C. (2012). UBY - A Large-Scale Unified Lexical- Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France.

Meyer, M. C. and Gurevych, I. (2012). *OntoWiktionary - Constructing an Ontology from the Collaborative Online Dictionary Wiktionary*. In Semi-Automatic Ontology Development: Processes and Resources, chapter 6, pp 131–161. IGI Global.

New, B. (2006). Lexique 3: Une nouvelle base de données lexicales. *Actes de la 13^e Conférence Traitement Automatique des Langues Naturelles (TALN 2006),* Louvain-la-Neuve, Belgium.

New, B., Brysbaert, M., Veronis, J. and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. In *Applied Psycholinguistics* (28): 661–677.

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 ACL Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. ACL-IJCNLP 2009, Singapore.

Pérennou, G. and De Calmès, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. In *Proceedings of ECST 1987*, pp. 1393–1396, Edinburgh, UK.

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

Romary, L., Salmon-Alt, S. and Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop on Electronic Dictionaries*, Geneva, Swiss.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., et Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S. (eds), *Advances in Natural Language Processing*, vol. 6233 of LNCS, 332–344. Springer.

Sajous, F., Hathout, N. and Calderone, B. (2013a). *GLÀFF, un Gros Lexique À tout Faire du Français*. Actes de la 20^e Conference sur le Traitement Automatique des Langues Naturelles (TALN 2013), pp. 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., Hathout, N. and Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Etudes et réalisations fondées sur le dictionnaire collaboratif. *Actes du 4^e Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, Germany.

Schultz, T., Vu, N. T. and Schlippe, T. (2013). GlobalPhone: A multilingual text & speech database in 20 languages. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*, pp. 8126–8130, Vancouver, Canada.

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

Storkel, H. L. and Hoover, J. R. (2011). The influence of part-word phonotactic probability/neighborhood density on word learning by preschool children varying in expressive vocabulary. In *Journal of Child Language*, 38, 628-643

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