# Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary

Nabil Hathout, Franck Sajous, Basilio Calderone

# Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary

**Nabil Hathout  Franck Sajous  Basilio Calderone**
CLLE-ERSS (CNRS & Université de Toulouse 2)

## Abstract

We present two approaches to automatically acquire morphologically related words from Wiktionary. Starting with related words explicitly mentioned in the dictionary, we propose a method based on orthographic similarity to detect new derived words from the entries' definitions with an overall accuracy of 93.5%. Using word pairs from the initial lexicon as patterns of formal analogies to filter new derived words enables us to rise the accuracy up to 99%, while extending the lexicon's size by 56%. In a last experiment, we show that it is possible to semantically type the morphological definitions, focusing on the detection of process nominals.

## 1 Introduction

Around the 1980s the computational exploitation of machine-readable dictionaries (MRDs) for the automatic acquisition of lexical and semantic information enjoyed a great favor in NLP (Calzolari et al., 1973; Chodorow et al., 1985). MRDs' definitions provided robust and structured knowledge from which semantic relations were automatically extracted for linguistic studies (Markowitz et al., 1986) and linguistic resources development (Calzolari, 1988). Today the scenario has changed as corpora have become the main source for semantic knowledge acquisition. However, dictionaries are regaining some interest thanks to the availability of public domain dictionaries, especially Wiktionary.

In the present work, we describe a method to create a morphosemantic and morphological French lexicon from Wiktionary's definitions. This type of large coverage resource is not available for almost all languages, with the exception of the CELEX database (Baayen et al., 1995) for English, German and Dutch, a paid resource distributed by the LDC.

The paper is organized as follows. Section 2 reports related work on semantic and morphological acquisition from MRDs. In Section 3, we describe how we converted Wiktionnaire, the French language edition of Wiktionary, into a structured XML-tagged MRD which contains, among other things, definitions and morphological relations. In Section 4, we explain how we used Wiktionnaire's morphological sections to create a lexicon of morphologically related words. The notion of morphological definitions and their automatic identification are introduced in Section 5. In Section 6, we show how these definitions enable us to acquire new derived words and enrich the initial lexicon. Finally, Section 7 describes an experiment where we semantically typed process nouns definitions.

## 2 Related work

Semantic relations are usually acquired using corpora (Curran and Moens, 2002; van der Plas and Bouma, 2005; Heylen et al., 2008) but may also be acquired from MRDs. MRDs-based approaches are bound to the availability of such resources. However, for some languages including French, no such resource exists. Recent years have seen the development of large resources built automatically by aggregating and/or translating data originating from different sources. For example, Sagot and Fišer (2008) have built WOLF, *"a free French Wordnet"* and Navigli and Ponzetto (2010) BabelNet, a large multilingual semantic network. Such resources tend to favor coverage over reliability and may contain errors and





*Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, pages 65–74,
Coling 2014, Dublin, Ireland, August 24 2014.

inaccuracy, or be incomplete. Pierrel (2013), while criticizing these resources, describes the digitization process of the *Trésor de la Langue Française*, a large printed French dictionary. The first impulse of this long-course reverse-engineering project is described in (Dendien, 1994) and resulted in the *TLFi*, a fine-grained XML-structured dictionary. Pierrel advocates mutualization, recommends resources sharing and underlines how the use of the TLFi would be relevant for NLP. Though we totally agree on this assertion, we deplore that the resource, being only available for manual use and not for download, prevents its use for NLP.

Crowdsourcing has recently renewed the field of lexical resources development. For example Lafourcade (2007) designed JeuxDeMots, a *game with a purpose*, to collect a great number of relations between words. Other works use the content of wikis produced by crowds of contributors. Initially in the shadow of Wikipedia, the use of Wiktionary tends to grow in NLP studies since its exploitation by Zesch et al. (2008). Its potential as an electronic lexicon was first studied by Navarro et al. (2009) for English and French. The authors leverage the dictionary to build a synonymy network and perform random walks to find missing links. Other works tackled data extraction: Anton Pérez et al. (2011) for instance, describe the integration of the Portuguese Wiktionary and Onto.PT; Sérasset (2012) built Dbnary, a multilingual network containing "easily extractable" entries. If the assessment of Wiktionary's quality from a lexicographic point of view has not been done yet, Zesch and Gurevych (2010) have shown that lexical resources built by crowds lead to results comparable to those obtained with resources designed by professionals, when used to compute semantic relatedness of words. In Sajous et al. (2013a), we created an inflectional and phonological lexicon from Wiktionary and showed that its quality is comparable to those of reference lexicons, while the coverage is much wider.

Comparatively little effort has been reported in literature on the exploitation of semantic relations to automatically identify morphological relations. Schone and Jurafsky (2000) learn morphology with a method based on semantic similarity extracted by latent semantic analysis. Baroni et al. (2002) combine orthographic (string edit distances) and semantic similarity (words' contextual information) in order to discover morphologically related words. Along the same line, Zweigenbaum and Grabar (2003) acquire semantic information from a medical corpus and use it to detect morphologically derived words. More recently, Hathout (2008) uses the TLFi to discover morphologically related words by combining orthographic and semantic similarity with formal analogy.

I another work, Pentheroudakis and Vanderwende (1993) present a method to automatically extract morphological relations from the definitions of MRDs. The authors automatically identify classes of morphologically related words by comparing the semantic information in the entry of the derivative with the information stored in the candidate base form. This effort shows the crucial importance and the potential of the MRDs' definitions to acquire and discover morphological relationships of derived words.

## 3 Turning the French Wiktionary into a Machine-Readable Dictionary

As mentioned is section 2, the quality of collaboratively constructed resources has already been assessed and we will not debate further the legitimacy of leveraging crowdsourced data for NLP purpose. We give below a brief description of Wiktionary[1] and of the process of converting it into a structured resource.

Wiktionary is divided in language editions. Each language edition is regularly released as a so-called *XML dump*.[2] The "XML" mention is somewhat misleading because it suggests that XML markups encode the articles' microstructure whereas only the macrostructure (articles' boundaries and titles) is marked by XML tags. Remaining information is encoded in *wikicode*, an underspecified format used by the *MediaWiki* content-management system. As explained by Sajous et al. (2013b) and Sérasset (2012), this loose encoding format makes it difficult to extract consistent data. One can choose to either restrict the extraction to prototypical articles or design a fine-grained parser that collects the maximum of the available information. The former goal is relatively easily feasible but leads to a resource containing only a small subset of Wiktionary's entries. Our belief is that the tedious engineering work of handling all

[1] For further details, read Zesch et al. (2008) and Sajous et al. (2013b).

[2] The dump used in this work is `https://dumps.wikimedia.org/frwiktionary/20140226/frwiktionary-20140226-pages-articles.xml.bz2`

```
== {{langue|fr}} ==
=== {{S|nom|fr}} ===
{{fr-rég|kurs}}
'''course''' {{pron|kurs|fr}} {{f}}
# [[action|Action]] de [[courir]], [[mouvement]] de celui qui [[court]].
#* ''[...], il n'est de bruit qu'un ver qui taraude incessamment les boiseries et dans le plafond,
la '''course''' d'un rongeur.'' {{source|{{w|Jean Rogissart}}, ''Passantes d'Octobre'', 1958}}
# {{sport|nocat=1}} Toute [[épreuve]] [[sportif|sportive]] où la [[vitesse]] est en jeu.
#* ''Nos pères étaient donc plus sages que nous lorsqu'ils repoussaient l'idée des '''courses'''.
# {{vieilli|fr}} [[actes|Actes]] d'[[hostilité]] que l'on faisait [[courir|en courant]] les mers
ou [[entrer|en entrant]] dans le [[pays]] [[ennemi]].
{{usage}} On dit maintenant [[incursion]], [[reconnaissance]], [[pointe]], etc.
#* ''Pendant les guerres de la révolution, Chausey, trop exposé aux '''courses''' des corsaires
de Jersey, resta inhabité.''
# {{figuré|fr}} [[marche|Marche]], [[progrès]] [[rapide]] d'une personne ou d'une chose.
#* ''Rien ne peut arrêter ce conquérant, ce fléau dans sa '''course'''.''

==== {{S|dérivés}} ====
* [[courser]]
* [[coursier]]
```

Figure 1: Wikicode extract of the noun *course*

wikicode particularities is valuable. In our case, it enabled us to design an unprecedented large copylefted lexicon that has no equivalent for French.

The basic unit of Wiktionary's articles is the word form: several words from different languages having the same word form occur in the same page (at the same URL). In such a page, a given language section may be divided in several parts of speech which may in turn split into several homonyms subsections. In the French Wiktionary, the *course* entry, for example, describes both the French and English lexemes. The French section splits into a noun section (*une course* 'a run; a race') and a section related to the inflected forms of the verb *courser* 'to pursue'. The noun section distinguishes 11 senses that all have definitions illustrated by examples. An extract of the noun section's wikicode is depicted in Figure 1. As can be seen, some wiki conventions are recurrent (e.g. double-brackets mark hyperlinks) and are easy to handle. Handling dynamic templates (marked by curly brackets) is more tricky. In definitions, they mark notes related to particular domains, registers, usages, geographic areas, languages, etc. In Figure 1, the pattern {{sport}} indicates that the second sense relates to the domain of sport ; the pattern {{vieilli|fr}} in the following definition denotes a dated usage ; the pattern {{figuré|fr}} in the last definition indicates a figurative one. We inventoried about 6,000 such templates and their aliases: for example, 4 patterns (abbreviated or full form, with or without ligature) signal the domain of enology: {{œnologie|fr}}, {{oenologie|fr}}, {{œnol|fr}} and {{oenol|fr}}. Unfortunately, the existence of such patterns does not prevent a contributor to directly write domain name in the page: several versions of "hardcoded domains" may be found, e.g. (oenologie) or (œnologie).

Inventorying all these variations enabled us: 1) to remove them from the definitions' text and 2) to mark them in a formal way. Thus, one can decide to remove or keep, on demand, entries that are marked as rare or dated, build a sublexicon of a given domain, remove diatopic variations or investigate only these forms (e.g. words that are used only in Quebec), etc.

The variations observed in the definitions also occur in phonemic transcriptions, inflectional features, semantic relations, etc. We focus here only on the information used in sections 6 and 7: definitions and morphological relations. However, we parsed Wiktionnaire's full content and extracted all kind of available information, handling the numerous variations that we observed to convert the online dictionary into a structured resource, that we called GLAWI.[3] It contains more than 1.4 million inflected forms (about 190,000 lemmas) with their definitions, examples, lexicosemantic relations and translations, derived terms and phonemic transcriptions. A shortened extract resulting from the conversion of the noun section of *course* is depicted in Figure 2. As can be seen, GLAWI includes both XML structured data and the initial corresponding wikicode. This version of the resource is intended to remain close to the Wiktionnaire's content, whereas other lexicons focused on a particular aspect will be released. Our aim is to provide ready-to-use lexicons resulting from different post-processing of GLAWI. Post-processing

[3]Resulting from the unification of GLÀFF and an updated version of WiktionaryX, GLAWI stands for *"GLÀFF and WiktionaryX"*. This resource is freely available at http://redac.univ-tlse2.fr/lexicons/glawi.html.

```
<pos type="nom" inlflected="0">
  <grammaticalInfo gender="f" number="s"/>
  <inflections>
    <infl form="courses" pos="Ncfp" lemma="course" prons="kuʁs"/>
  </inflections>
  <pron>kuʁs</pron>
  <definitions>
    <definition>
      <gloss>
        <txt>Action de courir, mouvement de celui qui court.</txt>
        <wiki>[[action|Action]] de [[courir]], [[mouvement]] de celui qui [[court]].</wiki>
      </gloss>
      <example>
        <wiki>''[…], il n'est de bruit qu'un ver qui taraude incessamment les boiseries et dans le plafond,
        la '''course''' d'un rongeur.'' {{source|{{w|Jean Rogissart}}, ''Passantes d'Octobre'', 1958}}</wiki>
        <txt>[…], il n'est de bruit qu'un ver qui taraude incessamment les boiseries et dans le plafond,
        la course d'un rongeur.</txt>
      </example>
    </definition>
    <definition>
      <gloss>
        <domain value="sport"/>
        <wiki>{{sport|nocat=1}} Toute [[épreuve]] [[sportif|sportive]] où la [[vitesse]] est en jeu.</wiki>
        <txt>Toute épreuve sportive où la vitesse est en jeu.</txt>
      </gloss>
      <example>
        <wiki>''Nos pères étaient donc plus sages que nous lorsqu'ils repoussaient l'idée des
        '''courses'''.'' {{source|J. Déhès, ''[[s:Essai sur l'amélioration des races chevalines de la
        France|Essai sur l'amélioration des races chevalines de la France]]'', 1868}}</wiki>
        <txt>Nos pères étaient donc plus sages que nous lorsqu'ils repoussaient l'idée des courses.</txt>
      </example>
    </definition>
    <subsection type="dérivés">
      <item>courser</item>
      <item>coursier</item>
    </subsection>
</pos>
```

Figure 2: Extract of the noun subsection of *course* converted into a workable format

steps will consist in 1) selecting information relevant to a particular need (e.g. phonemic transcriptions, semantic relations, etc.) and 2) detecting inconsistencies and correcting them. The initial GLAWI resource, containing all the initial information, will also be released so that anyone can apply additional post-processings. GLAWI unburdens such users from the efforts of parsing the wikicode.

Articles from Wiktionnaire may contain morphologically derived terms. Figures 1 and 2 show that *course* produces the derived verb *courser* and noun *coursier* 'courier'. Such derivational relations are collected from Wiktionnaire and included in GLAWI. We show below how we leverage this information, in addition to GLAWI's definitions, to acquire morphological and morphosemantic knowledge.

## 4 Acquisition of morphological relations from GLAWI morphological subsections

We first extracted from GLAWI the list of the lexeme headwords that have typographically simple written forms (only letters) and that belong to the major POS: noun, verb, adjective, and adverb. This list (GLAWI-HW) contains 152,567 entries: 79,961 nouns, 22,646 verbs, 47,181 adjective and 2,779 adverbs). In what follows, we only consider these words.

Then we created a morphological lexicon extracted from the morphological subsections[4] of GLAWI (hereafter GMS). The lexicon consists of all pairs of words $(w_1, w_2)$, where $w_1$ and $w_2$ belong to GLAWI-HW and where $w_2$ is listed in one of the morphological subsections of the article of $w_1$ or vice versa. GMS contains 97,058 pairs. The extraction of this lexicon from GLAWI was very simple, all the variability in Wiktionnaire's lexicographic descriptions being supported by our parser (see Section 3).

The remainder of the paper presents two methods for extending GMS. In a first experiment, we complement this lexicon with new pairs acquired from GLAWI's definitions. In a second one, we show how some of GMS's morphological pairs can be classified with respect to a given semantic class.

[4]The morphological subsections appear under 4 headings in Wiktionnaire: *apparentés*; *apparentés étymologiques*; *composés*; *dérivés*.

| $w_1$ | $w_2$ | $w_1$ | $w_2$ |
|---|---|---|---|
| bisannuel_A | an_N | républicain_N | république_N |
| compilation_N | compilateur_A | similaire_A | dissimilitude_N |
| foudroyeur_A | foudre_N | tabasser_V | tabassage_N |
| militance_N | militer_V | taxidermie_N | taxidermiser_V |
| presse_N | pression_N | volcan_N | volcanique_A |

Figure 3: Excerpt of GMS lexicon. Letters following the underscore indicate the grammatical category.

## 5 Morphological definitions

Basically, a dictionary definition is a pair composed of a word and a gloss of its meaning. In the following, we will use the terms **definiendum** for the defined word, **definiens** for the defining gloss and the notation *definiendum* = *definiens*. The definition articulates a number of lexical semantic relations between the definiendum and some words of the definiens as in (1) where *chair* is a hyponym of *furniture*, is the holonym of *seat*, *legs*, *back* and *arm rests* and is also the typical instrument of *sit on*. Some of the relations are made explicit by lexical markers as *used to* or *comprising*.

(1) chair$_N$ = An item of **furniture** used to **sit on** or in comprising a **seat**, **legs**, **back**, and sometimes **arm rests**, for use by one person.

Martin (1983) uses these relations to characterize the definitions. In his typology, definitions as in (2) are considered to be (morphological) derivational because the definiendum is defined with respect to a morphologically related word. In these definitions, the lexical semantic relation only involves two words that are morphologically related. Being members of the same derivational family, the orthographic representations of these words show some degree of similarity that can help us identify the morphological definitions. In (2) for example, the written forms *nitrificateur* 'nitrifying' and *nitrification* 'nitrification' share a 10 letters prefix and only differ by 3 letters. This strong similarity is a reliable indicator of their morphologically relatedness (Hathout, 2011b). Building on this observation, a definition is likely to be morphological if its definiens contains a word which is orthographically similar to the definiendum.

(2) nitrificateur$_A$ = Qui produit, qui favorise la **nitrification**.
'nitrifying' 'that produces, that favors nitrification'

We used Proxinette, a measure of morphological similarity defined in (Hathout, 2008), to identify the morphological definitions. Proxinette is designed to reduce the search space for derivational analogies. The reduction is obtained by bringing closer the words that belong to the same derivational families and series, since it is precisely within these paradigms that an entry is likely to form analogies (Hathout, 2011a). Proxinette describes the lexemes by all the $n$-grams of characters that appear in their inflected forms in order to catch the inflectional stem allomorphy because it tends to also show up in derivation (Bonami et al., 2009). The $n$-grams have an additional tag that indicates if they occur at the beginning, at the end or in the middle of the word. This information is described by adding a `#` at the beginning and end of the written forms. For example, in Figure 4, *localisation* 'localization', *localiser* 'localize; locate' and *focalisation* 'focalization' share the `ions#` ending because it occurs in their inflected forms *localisations* (plural), *localisions* (1st person plural, indicative, imperfect) and *focalisations* (plural). $n$-grams of size 1 and 2 are ignored because they occur in too many words and are not discriminant enough. Proxinette builds a bipartite graph with the words of the lexicon on one side and the features ($n$-grams) that characterize them on the other. Each word is linked to all its features and each feature is connected to the words that own it (see Figure 4). The graph is weighted so that the sum of weights of the outgoing edges of each node is equal to 1. Morphological similarity is estimated by simulating the spreading of an activation. For a given entry, an activation is initiated at the node that represents it. This activation is then propagated towards the features of the entry. In a second step, the activations in the feature nodes are propagated towards the words that possess them. The words which obtain the highest activations are the most similar to the entry. The edge weights and the way the graph is traversed brings closer the words that share the largest number of common features and the most specific ones (i.e. the less frequent).

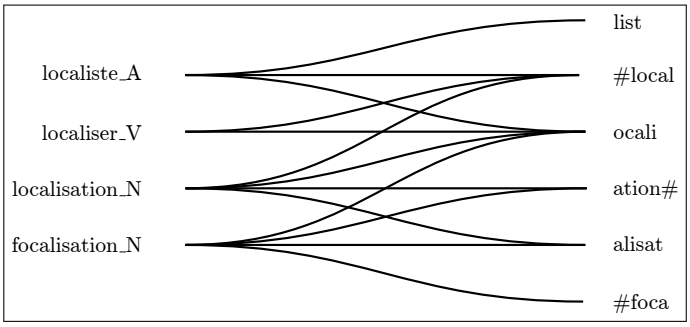


Figure 4: Excerpt of Proxinette bipartite graph. The graph is symmetric.

**écholocalisation**_N **relocalisation**_N **radiolocalisation**_N **géolocalisation**_N glocalisation_N **délocalisation**_N **antidélocalisation**_A **localisateur**_N **localisateur**_A vocalisation_N focalisation_N **localiser**_V **localisable**_A **délocalisateur**_N **localisé**_A **localiste**_N **localiste**_A **localisme**_N tropicalisation_N

Figure 5: The most similar words to the noun *localisation*. Words in boldface belong to the derivational family of *localisation*. Words in light type belong to its derivational series.

We applied Proxinette to GLAWI-HW and calculated for each of them a neighborhood consisting of the 100 most similar words. Figure 5 shows an excerpt of the neighborhood of the noun *localisation*. The occurrence of the verb *localiser* in this list enables us to identify the morphological definition (3).

(3) localisation$_{N}$ = Action de **localiser**, de se **localiser**.
'localization' 'the act of localizing, of locating'

The two experiments we conducted use the same data, namely the morphological definitions of GLAWI. These definitions are selected as follows:

1. We extracted all GLAWI definition glosses (definientia) with their entries and POS (definienda).

2. We syntactically parsed the definientia with the Talismane dependency parser (Urieli, 2013). Figure 6 presents the dependencies syntactic trees for the definientia in (4).

3. We tagged as morphological all definitions where, in the parsed definiens, at least one lemma (henceforth referred to as morphosemantic head) occurs in the definiendum neighborhood. For example, in (4), both definitions are tagged as morphological because *arrêter* occurs in the neighborhood of *arrêt*, and *découronner* and *couronne* occur in that of *découronnement*.

(4) a. arrêt$_{N}$ = Action de la main pour arrêter le cheval.
'stop' 'action of the hand to stop the horse'

b. découronnement$_{N}$ = L'action de découronner, d'enlever la couronne.
'uncrowning' 'the act of uncrowning, of removing the crown'

Morphosemantic heads may be the derivational base of the definiendum like *découronner*, a more distant ancestor like *couronne* or a "sibling" like in (2) where *nitrification* is a derivative of the definiendum base *nitrifier* 'nitrify'.

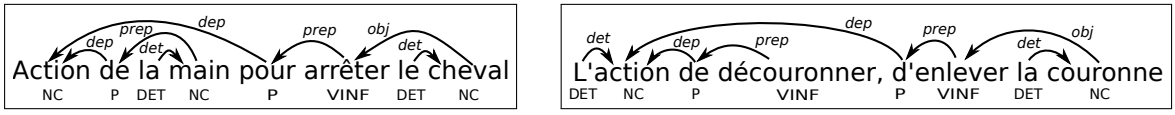


Figure 6: POS-tags and syntactic dependencies of the definientia of (4).

## 6 Acquisition of morphological relations from GLAWI morphological definitions

We extracted from GLAWI's morphological definitions the pairs of words $(w_1, w_2)$ where $w_1$ is the definiendum and $w_2$ the definiens morphosemantic heads (or one of its morphosemantic head if it has many). After symmetrization, we obtained a lexicon (hereafter GMD) of 107,628 pairs. 32,256 of them belongs to GMS. A manual check of the 75,372 remaining pairs would enable its addition to GMS.

GMD additional pairs have been evaluated by three judges in two steps. The judges were instructed to set aside the orthographic variants as *desperado_N* / *despérado_N*. We first randomly selected 100 pairs and had them checked by three judges in order to estimate the inter-annotator agreement. The average F-measure of the agreement is 0.97 ; Fleiss's kappa is 0.65. The judges then checked 100 randomly selected pairs each. 9 out of the 300 pairs were variants and 19 errors were found in the 291 remaining ones which results in an overall accuracy of 93.5%. This method would lead to an increase of GMS by more than 70,000 pairs.

The general quality of these acquired pairs can be significantly increased by formal analogy filtering. The idea is to use analogy as a proxy to find pairs of words that are in the same morphological relation. GMS pairs being provided by Wiktionary contributors, we consider them as correct and use them as analogical patterns to filter out the pairs acquired from the morphological definitions. By formal analogy, we mean an analogy between the orthographic representations. For instance, the GMD pair *citrique_A*:*citron_N* form an analogy with *électrique_A*:*électron_N*. The latter being correct, we can assume that the former is correct too.

(5) a. citrique_A : citron_N = électrique_A : électron_N

b. fragmentation_N : défragmenter_V = concentration_N : déconcentrer_V

Analogies between strings are called formal analogies (Lepage, 2003; Stroppa and Yvon, 2005). One way to check a formal analogy is to find a decomposition (or factorization) of the four strings such that the differences between the first two are identical to the ones between the second two. In the analogy in (5a), the ending `ique` is replaced by `on` and the POS `A` by `N` in both pairs. We applied analogical filtering to GMS and GMD pairs. 86,228 pairs in GMD form at least one analogy with a pair in GMS; 53,972 of them do not occur in GMS. 300 of these pairs have been checked by three judges. They only found 3 variants and one error. The obtained accuracy is therefore over 99% (see Table 1).[5]

| | initial | | analogical | |
|---|---|---|---|---|
| | pairs | accuracy | pairs | accuracy |
| GMS | 97,058 | – | – | – |
| GMD | 107,628 | 95.4% | 86,228 | 99.8% |
| GMD \ GMS | 75,372 | 93.5% | 53,972 | 99.7% |

Table 1: Summary of the quantitative results

GMD morphological relations will not be included into GLAWI. GMS and GMD are made available as separate resources on the GLAWI web page.

## 7 Semantic typing of the morphological definitions

The next experiment aims to demonstrate that morphological definitions could easily and quite accurately be typed semantically. We focus on a particular semantic type, namely definitions of process nominals such as (6) because they can be evaluated with respect to the Verbaction database (Hathout and Tanguy, 2002). Deverbal nominals have been extensively studied in linguistics (Pustejovsky, 1995) and used in a number of tools for various tasks. One of their distinctive feature is that they almost have the same meaning as their base verb. For instance, in (7) the noun and verb phrases are paraphrases of one another. Verbaction contains 9,393 verb-noun pairs where the noun is morphologically related to the verb and can be used to express the act denoted by the verb (e.g. *verrouiller*:*verrouillage*).[6] It has been

[5]Unfortunately, these results could not have been compared with those of Pentheroudakis and Vanderwende (1993) because their system makes use of a number of lexical and semantic resources that are not available for French. However, a comparison with Baroni et al. (2002) is underway although their method is corpus-based (and not MRD-based).

[6]Verbaction is freely available at: `http://redac.univ-tlse2.fr/lexiques/verbaction.html`.

used in syntactic dependency parsing by Bourigault (2007), in the construction of the French TimeBank by Bittar et al. (2011), in question answering systems by Bernhard et al. (2011), etc.

(6) verrouillage$_{N}$ = Action de verrouiller.
'locking' 'the act of locking.'

(7) nous **vérouillons la porte** rapidement 'we quickly lock the gate'
le **verrouillage de la porte** est rapide 'gate locking is quick'

In our experiment, we used the linear SVM classifier *liblinear* of Fan et al. (2008) to assign a semantic type to the definitions that have a nominal definiendum and where the morphosemantic head of the definiens is a verb as in (4) or (6). Verbaction was used to select a corpus of 1,198 of such definitions. Three judges annotated them. 608 definientia were tagged as processive and 590 ones as non processive. We then divided the corpus into a test set made up of 100 processive and 100 non processive definitions and a training set consisting of the remaining definientia.

The classifier is trained to recognize that the definientia in (4) express the same semantic relation between the morphosemantic head of the definiens and the definiendum. We use the method proposed by (Hathout, 2008) to capture this semantic similarity. Definientia are described by a large number of redundant features based on lemmata, POSs and syntactic dependencies. The features are $n$-grams calculated from Talismane parses (see figure 6). They are defined as follows:

1. We first collect all the paths that go from one word in the definiens to the syntactic root (e.g. [*arrêter*, *pour*, *action*] is a path that starts at *arrêter* in (4a)).

2. We extract all the $n$-grams of consecutive nodes in these paths.

3. Each $n$-gram yields 3 features: the sequence of the node's lemmata, the sequence of the nodes POS, and the sequence of syntactic dependency relations.

We obtained an accuracy of 97% for the semantic typing of the 200 definientia of the test set. The most immediate application of the classifier is the enrichment of Verbaction. Running the classifier on all the definitions with a nominal definiendum and a verbal morphosemantic head will provide us with new couples that could be added to the database. The classifier could also help us type process nouns that are not morphologically derived such as *audition* 'hearing' which is defined with respect to the verb *entendre* 'hear'. Similar typing could be performed for other semantic types such as agent nouns (in *-eur* or *-ant*), change of state verbs (in *-iser* or *-ifier*) or adjectives expressing possibility (in *-able*), etc. The experiment also shows that morphological definitions are well suited for semantic analysis because they express regular semantic relationship between pairs of words that are distinguished by their orthographic similarity.

## 8 Conclusion

In this paper, we have presented GLAWI, an XML machine-readable dictionary created from Wiktionnaire, the French edition of the Wiktionary project. We then showed that GLAWI was well suited for conducting computational morphology experiments. GLAWI contains morphological subsections which provide a significant number of valid and varied morphological relations. In addition, morphological relations can also be acquired from GLAWI morphological definitions. We presented a method to identify these definitions and the words in relation with a fairly good accuracy. We then used formal analogy to filter out almost all the erroneous pairs acquired from morphological definitions. In a second experiment, we demonstrate how to assign the morphological definitions to semantic types with a high accuracy.

This work opens several research avenues leading to a formal representation of the different form and meaning relations that underlie derivational morphology. The next move will be to organize the morphological relations into a graph similar to Démonette (Hathout and Namer, 2014) and identify the paradigms which structure them. We also plan to apply the semantic classification to other semantic types which could ultimately enable us to explore the intricate interplay between form and meaning.

## References


Leticia Anton Pérez, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pages 703–717, Lisbon, Portugal.

Rolf Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, Philadelphia, PA.

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-2002*, pages 48–57, Philadelphia, PA, USA.

Delphine Bernhard, Bruno Cartoni, and Delphine Tribout. 2011. A Task-Based Evaluation of French Morphological Resources and Tools. *Linguistic Issues in Language Technology*, 5(2).

André Bittar, Pascal Amsili, Pascal Denis, et al. 2011. French TimeBank: un corpus de référence sur la temporalité en français. In *Actes de la 18e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2011)*, volume 1, pages 259–270, Montpellier, France.

Olivier Bonami, Gilles Boyé, and Françoise Kerleroux. 2009. L'allomorphie radicale et la relation flexion-construction. In Bernard Fradin, Françoise Kerleroux, and Marc Plénat, editors, *Aperçus de morphologie du français*, pages 103–125. Presses universitaires de Vincennes, Saint-Denis.

Didier Bourigault. 2007. *Un analyseur syntaxique opérationnel : SYNTEX*. Habilitation à diriger des recherches, Université Toulouse II-Le Mirail, Toulouse.

Nicoletta Calzolari, Laura Pecchia, and Antonio Zampolli. 1973. Working on the Italian Machine Dictionary: A Semantic Approach. In *Proceedings of the 5th Conference on Computational Linguistics - Volume 2*, pages 49–52, Stroudsburg, PA, USA.

Nicoletta Calzolari. 1988. The dictionary and the thesaurus can be combined. In Martha Evens, editor, *Relational Models of the Lexicon*, pages 75–96. Cambridge University Press.

Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, ACL '85, pages 299–304, Stroudsburg, PA, USA.

James R. Curran and Marc Moens. 2002. Improvements in Automatic Thesaurus Extraction. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, USA.

Jacques Dendien. 1994. Le projet d'informatisation du TLF. In Éveline Martin, editor, *Les textes et l'informatique*, chapter 3, pages 31–63. Didier Érudition, Paris, France.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Nabil Hathout and Fiammetta Namer. 2014. La base lexicale démonette : entre sémantique constructionnelle et morphologie dérivationnelle. In *Actes de la 21e conférence annuelle sur le traitement automatique des langues naturelles (TALN-2014)*, Marseille, France.

Nabil Hathout and Ludovic Tanguy. 2002. Webaffix : Finding and validating morphological links on the WWW. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1799–1804, Las Palmas de Gran Canaria, Spain.

Nabil Hathout. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the Coling workshop Textgraphs-3*, pages 1–8, Manchester, England.

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

Nabil Hathout. 2011b. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In *Des unités morphologiques au lexique* (Roché et al., 2011), pages 251–318.

Kris Heylen, Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. 2008. Modelling Word Similarity: an Evaluation of Automatic Synonymy Extraction Algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Mathieu Lafourcade. 2007. Making People Play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, Pattaya, Thailand.

Yves Lepage. 2003. *De l'analogie rendant compte de la commutation en linguistique*. Habilitation à diriger des recherches, Université Joseph Fourier, Grenoble.

Judith Markowitz, Thomas Ahlswede, and Martha Evens. 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, pages 112–119, Stroudsburg, PA, USA.

Robert Martin. 1983. *Pour une logique du sens*. Linguistique nouvelle. Presses universitaires de France, Paris.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Ivy Kuo, Pierre Magistry, and Chu-Ren Huang. 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.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of ACL'2010*, pages 216–225, Uppsala, Sweden.

Joseph Pentheroudakis and Lucy Vanderwende. 1993. Automatically identifying morphological relations in machine-readable dictionaries. In *Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research*, pages 114–131.

Jean-Marie Pierrel. 2013. Structuration et usage de ressources lexicales institutionnelles sur le français. *Linguisticae investigationes Supplementa*, pages 119–152.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Michel Roché, Gilles Boyé, Nabil Hathout, Stéphanie Lignon, and Marc Plénat. 2011. *Des unités morphologiques au lexique*. Hermès Science-Lavoisier, Paris.

Benoît Sagot and Darja Fišer. 2008. Building a Free French Wordnet from Multilingual Resources. In *Proceedings of OntoLex 2008*, Marrakech, Morocco.

Franck Sajous, Nabil Hathout, and Basilio Calderone. 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.

Franck Sajous, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, and Yannick Chudy. 2013b. Semi-automatic enrichment of crowdsourced synonymy networks: the WISIGOTH system applied to Wiktionary. *Language Resources and Evaluation*, 47(1):63–96.

Patrick Schone and Daniel S. Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, pages 67–72, Lisbon, Portugal.

Gilles Sérasset. 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, Turkey.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Procs. of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 120–127, Ann Arbor, MI.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse-Le Mirail.

Lonneke van der Plas and Gosse Bouma. 2005. Syntactic Contexts for Finding Semantically Related Words. In Ton van der Wouden, Michaela Poß, Hilke Reckman, and Crit Cremers, editors, *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*, volume 4 of *LOT Occasional Series*. Utrecht University.

Torsten Zesch and Iryna Gurevych. 2010. Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering.*, 16(01):25–59.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Pierre Zweigenbaum and Natalia Grabar. 2003. Learning derived words from medical corpora. In *9th Conference on Artificial Intelligence in Medicine Europe*, pages 189–198.