Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3877–3885
Marseille, 11–16 May 2020

Glawinette : a linguistically motivated derivational description of French acquired from GLAWI

Nabil Hathout[1], Franck Sajous[1], Basilio Calderone[1], Fiammetta Namer[2]
[1]CLLE, CNRS & Université de Toulouse Jean Jaurès, [2]ATILF, Université de Lorraine & CNRS

Abstract

Glawinette is a derivational lexicon of French that will be used to feed the Démonette database. It has been created from the GLAWI machine readable dictionary. We collected pairs of words from the definitions and the morphological sections of the dictionary and then selected the ones that form regular formal analogies and that instantiate frequent enough formal patterns. The graph structure of the morphological families has then been used to identify for each pair of lexemes derivational patterns that are close to the intuition of the morphologists.

Keywords: derivational database, French, paradigmatic morphology, derivational family, derivational series

1. Introduction

The aim of this work is to create a derivational lexicon of French that could be used to feed Démonette, a large coverage morphological database which combines the results of various linguistic studies (Hathout and Namer, 2014a; Hathout and Namer, 2014b; Hathout and Namer, 2016; Namer et al., 2019). The idea behind the creation of this lexicon, named Glawinette, is to take advantage of the availability of GLAWI (Sajous and Hathout, 2015; Hathout and Sajous, 2016), a large French machine readable dictionary derived from Wiktionnaire, the French edition of Wiktionary. The task addressed in this article is to discover the derivational relations that hold in a subset of the French lexicon. Because these relations cannot be identified only from the formal properties of words (Hathout, 2002), some semantic knowledge must be used in order for instance to exclude the numerous cases where the formal relation between a pair of morphologically unrelated words (such as *poisse*:*poisson* 'unluck:fish') is the same as the relation between morphologically related words such as *corde*:*cordon* 'rope:string', *poche*:*pochon* 'bag:small bag', *glace*:*glaçon* 'ice:ice cube', etc.

Lexicographic definitions provide precise descriptions of word meanings that could help us perform more accurate morphological analysis (Hathout, 2009a; Hathout, 2011a; Hathout et al., 2014). However, we cannot automatically extract from the definitions all the morphosemantic properties of the derivational relations between derived headwords and their bases. For instance, we cannot abstract from the definitions of *-on* suffixed words such as in (1) that these words denote an entity that belongs to the category of the base and that is small in this category (Plénat, 2005).

(1) a. *clocheton*: *petit bâtiment en forme de clocher, de tourelle, dont on orne les angles ou le sommet d'une construction* 'small building in the shape of a bell tower, that decorate buildings corners or tops'

b. *glaçon*: *morceau de glace* 'piece of ice'

Yet, the definitions of derived words usually provide a very important information, namely the (semantic) base of these headwords. For instance, the definition in (1a) includes *clocher*, the base of *clocheton*. More generally, derived words are usually defined by means of "morphological definitions" (Martin, 1992), that is definitions that relate the headword to a member of its derivational family. The work we present is based on this observation. We also take advantage of the paradigmatic organization of derivational morphology (Van Marle, 1985; Bauer, 1997; Hathout and Namer, 2018b; Hathout and Namer, 2019; Bonami and Strnadová, 2019) to identify a number of additional characteristics of the derivatives and their relations, and in particular to identify their derivational series, families and exponents or affixes (Hathout, 2009b; Hathout, 2011b).

In addition to the definitions, GLAWI directly provides derivational relations in the morphological sections (derivatives, related words) and derivational descriptions in the etymological sections. We used the former for the creation of Glawinette but not the latter because etymological sections are messy, hard to parse and sometimes inaccurate.

The remainder of the article is organized as follows: Section 2. presents the resource and the two ways in which morphological relations can be organized. Section 3. puts our work in a broader context and connects it to similar studies. We then describe in Section 4. how we extracted candidate pairs from the definitions and in Section 5. how we collected additional pairs from the morphological sections of GLAWI. These candidates are then screened in order to select the ones that are connected by valid derivational relations. Section 6. details the selection of the correct pairs by means of formal analogies and, Section 7., by means of patterns that describe formal regularities in the series of pairs.

2. Two projections of the French derivational lexicon

Glawinette contains 97,293 words connected by 47,712 relations. Its design has benefited from the recent debates on the nature of derivational morphology and from the proposals that it is paradigmatic in nature (Hathout and Namer, 2018b; Bonami and Strnadová, 2019; Hathout and Namer, 2018a; Hathout and Namer, 2019). Its design also benefited from the work done on Démonette (Hathout and Namer, 2014a; Hathout and Namer, 2014b; Hathout and Namer, 2016).

In Morphonette, Hathout (2011a) proposes to represent the morphological relations by filaments, *i.e.*, series of words

involved in identical derivational relations. In Démonette, Hathout and Namer (2014a) and Namer et al. (2019) propose a redundant representation where all the derivational relations that exist in the lexicon are listed, be they direct or indirect. The other dimensions of the morphological organization (*i.e.*, the families and the series) are reconstructed from the graph of relations (for the derivational families), from the exponents (for the morphological series) and from the abstract definitions (for the semantic series).

Glawinette combines and extends these two manners to describe the derivational structure of the lexicon. It gives two complementary projections of the morphological organization, namely (*i*) all the derivational families defined by the derivational relations and (*ii*) all the derivational series. Derivational families are connected graphs of derivational relations. These graphs can be described by listing the set of their edges, that is the derivational relations they include. The derivational families of *prince* 'prince' and *introduire* 'introduce' are given in Figure 1. In the first, we can see that the relations are symmetrical (*e.g.*, the family includes *prince*:*princesse* 'princess' and *princesse*:*prince*), some are direct, like *prince*:*princier* 'princely'$_{\text{Adj}}$, the latter being the relational adjective of *prince* and other are indirect like *introducteur*:*introduction* 'introducer:introduction' in the family of *introduire*, where both nouns are derived from the verb *introduire*. Because most of these relations are extracted from definitions, the families are usually not complete graphs. For instance, in the family of *prince*, *princesse* is not connected to *princier*. Moreover, some relations are not symmetrical. Glawinette contains 15,904 families, the largest having 340 relations (the family of *forme* 'form') and the smallest, only one (*e.g.*, *édeniser*:*éden* 'Edenize:Eden').

```
prince=N:princesse=N
prince=N:princier=A                    'princely'Adj
prince=N:princillon=N                  'petty prince'
prince=N:princiser=V                   'make become a prince'
princesse=N:prince=N
princier=A:prince=N
princier=A:princièrement=R             'princely'Adv
princillon=N:prince=N
princiser=V:prince=N
princièrement=R:princier=A
```

```
introducteur=N:introduction=N
introducteur=N:introduire=V
introductif=A:introduire=V             'introductive'
introduction=N:introducteur=N
introduction=N:introductoire=A         'introductory'
introduction=N:introduire=V
introduction=N:réintroduction=N        'reintroduce'
introductoire=A:introduction=N
introduire=V:introducteur=N
introduire=V:introductif=A
introduire=V:introduction=N
réintroduction=N:introduction=N
```

Figure 1: The derivational families of *prince* 'prince' and *introduire* 'introduce' in Glawinette

The second projection Glawinette provides is the derivational series of the lexicon. A derivational series is a set of pairs of lexemes connected by the same relation and, therefore that instantiate the same formal patterns. For instance, the series that contains the *formateur*:*formation* 'trainer:training' is made up of pairs of nouns where the first ends in *-eur*, the second in *-ion* and where the stem that precedes *-eur* in the first and *-ion* in the second are identical. This common sequence is for instance `format` in *formateur*:*formation*. In Glawinette, these regularities are described by patterns made up of a regular expression and a grammatical category. For instance, the patterns of the series of *formateur*:*formation* is `^(.+)eur$=N:^(.+)ion$=N` where `(.+)` represents the initial sequence shared by the pair of words. Table 1 presents an excerpt of the series of this pair.

`^(.+)eur$=N`	`^(.+)ion$=N`
acteur	action
animateur	animation
classificateur	classification
colonisateur	colonisation
directeur	direction
décentralisateur	décentralisation
dépresseur	dépression
éditeur	édition
expositeur	exposition
formateur	formation
réacteur	réaction
réviseur	révision

Table 1: Excerpt of the derivational series of *formateur*:*formation*

One contribution of Glawinette is to provide linguistically motivated exponents for the derivational series and consequently for the series of words. Figure 2 illustrates this distinctive characteristics for an excerpt of the family of *forme*. As we can see, the patterns are not obtained by simply removing the largest common substring in the two words. For instance, *formalisme*:*formaliser* 'formalism:formalize' is characterized by the exponents *-isme* and *-iser* and not by the most general patterns `^(.+)me$` and `^(.+)er$`. Glawinette contains 5400 series of relations, the largest being the noun-verb conversion (*i.e.*, zero-derivation) `^(.+)e$=N:^(.+)er$=V` with 3918 pairs. The smallest series contain pairs with idiosyncratic written forms such as *naïf*:*naïve* 'naive$_{\text{Masc}}$:naive$_{\text{Fem}}$'.

formalisme:*formaliser*	`^(.+)isme$=N`	`^(.+)iser$=V`
formatif:*formation*	`^(.+)atif$=A`	`^(.+)ation$=N`
formellement:*formel*	`^(.+)ellement$=R`	`^(.+)el$=A`
réforme:*réformette*	`^(.+)e$=N`	`^(.+)ette$=N`
réformiste:*réformisme*	`^(.+)iste$=A`	`^(.+)isme$=N`

Figure 2: Series are characterized by linguistically motivated exponents. The English equivalent of the four last examples are 'formative:formation', 'formally:formal', 'reform:small reform', 'reformist:reformism'

3. Related work

The supply in morphological resources has increased significantly in recent years, for inflection and derivation. Inflectional lexicons have become available for many languages because they are required in important NLP tasks such as syntactic parsing, spell checking or to develop predictive keyboards. On the other hand, modern NLP systems do not make use of derivational resources. Often, derivation is summarily taken into account and dealt with at the level of tokenization as in FastText (Bojanowski et al., 2017) or BERT (Devlin et al., 2018). Derivational databases are mainly used in psycho-linguistics to create experimental material, in speech and language remediation to setup exercises or to teach vocabulary at primary schools.

One way, derivational databases can be created is by running a morphological analyzer on a lexicon or a large corpus. Many of these systems are based on machine learning and trained to decompose words into morphemes, like Linguistica (Goldsmith, 2001), Morfessor (Creutz and Lagus, 2005) or, more recently, the model of (Cotterell and Schütze, 2018); for a panorama, see (Bernhard et al., 2011).

The other way to create derivational databases is to annotate lexicons semi-automatically. This annotation can take multiple forms: in CELEX (Baayen et al., 1995), the words are decomposed into morphemes; DeriNet (Vidra et al., 2019) describes the morphological relations between derived words and their bases; in CATVAR (Habash and Dorr, 2003) or DErivBase (Zeller et al., 2013) the lexicon is clustered into derivational families; for an exhaustive presentation of the derivational databases available for Romance, Germanic and Slavic languages, see (Kyjánek, 2018). The main contributions of Glawinette with respect to these resources are (*i*) the combination of the derivational families with the derivational series, and (*ii*) the description of the morphological relations by means of linguistically relevant exponents.

For many year, French lacked a large-scale derivational resources similar to CELEX. The first derivational databases aimed at closing the gap is Démonette (Hathout and Namer, 2014a; Hathout and Namer, 2016). Its first version describes 7542 derivational families made up of a verb and its agent and action nouns and modality adjective. The development of Démonette has three objectives: (*i*) produce a resource whose entries are derivational relations between pairs of words labeled with linguistically grounded features; the description is provided by DériF's analyses (Namer, 2009; Namer, 2013) and includes a semantic description of the relation; (*ii*) complement these $W_1 \rightarrow W_2$ derivations by relations that hold in the derivational families provided by Morphonette (Hathout, 2009a); (*iii*) define an extensible and redundant architecture, which can be fed by varied and heterogeneous morphological resources. The design of Glawinette makes its content appropriate to feed Démonette, by means of new entries and new relations that complement its current families.

4. Extraction of the derivational relations present in the definitions

Glawinette is created in three steps. We first collect pairs of potentially related words from GLAWI's morphological sections and definitions. We then select the pairs for which we are confident that they are correct. The third step is to provide these pairs of lexemes with patterns that describe linguistically motivated exponents (affixes).

As we already mentioned in the introduction, most derived words are defined with respect to another lexeme of their derivational family. This lexeme is usually the base but could be another word. For instance, *développer* 'to develop' is the base of *développement* 'development' in (2a) while in (2b), *productivisme* 'productivism' is defined with respect to *production* 'production' which is clearly not its base. However, it is also defined with respect to *productivité* 'productivity') in the second part of the definition, which could be considered as the (semantic) base of *productivisme* even if formally, both are constructed on the adjective *productif* 'productive'.

(2) a. *développement = action de développer, de se développer ou résultat de cette action, au propre et au figuré* 'action of developing or result of this action, both literally and figuratively'

b. *productivisme = doctrine selon laquelle la production est un objectif premier, système qui prône le sacrifice de toute autre considération pour maximiser la productivité* 'doctrine which states that production is a primary objective, system that advocates the sacrifice of all other considerations to maximize productivity'.

The extraction of derivational relations from the definitions is circular, to some extent. In order to find the morphological relations, we must know which words are derivatives but this is precisely the information we want to extract from the dictionary. One effective solution to this problem could be to use the Proxinette measure (Hathout, 2008; Hathout, 2014) in order to identify the possible members of their derivational families and series. The idea would be to only consider the very first neighbors of the headword that occur in the definition. However, some derivatives may be missed because Proxinette is not categorical. Moreover, we would still have to exclude the pairs of words where the neighborhoods returned by Proxinette belong to the derivational series of the entry.

We therefore opted for a different and more simple solution. We made use of the hypothesis that derivational relations connect forms that display regular alternations. More precisely, we have used the syntactically parsed version of the definitions and considered all the pairs made up of the headword and any noun, verb, adjective or adverb that occur in one of its definitions. The procedure yields 2,184,847 pairs of words.

5. Collecting the derivational relations described in the morphological sections

The second source of candidate pairs are the GLAWI's morphological sections. GLAWI being fully normalized and structured, it is very easy to collect the morphological relations described in the morphological section. These relations are of three types: derivatives, compounds and related words.

- Derivatives are words derived from the entry such as *motorisation* 'motorization' in the article of *motoriser* 'motorize'.
- Compounds are words that have the entry as one of their components such as *anticolonialisme* 'anticolonialism' in the article of *colonialisme* 'colonialism'.
- Related words are words belonging to the same derivational family that are not directly coined from the entry through derivation or compounding like *colon* 'settler' in the article of *colonialisme*.

In practice, the boundaries between these three types are relatively blurred in Wiktionnaire: compounds may be described as derivatives (*stéréocomparateur* 'stereo comparator' in the article of *comparateur* 'comparator') and vice-versa; derivatives may be considered as related words (*distorsion* 'distortion' in the article of *distordre* 'distort'), etc. Since our aim is to collect valid derivational relations from GLAWI and given that all the information present in Wiktionnaire has (in theory) been added by humans with a fair command of French, we could include all the information present in these sections into Glawinette. However, Wiktionnaire's contributors are not morphologists and do not all have the same conception of which relation is morphological or not. For instance, some include etymologically related lexemes while others only consider synchronic relations. Some consider domain specific usages while others limit the morphological descriptions to the relations that hold in general language, etc. For these reasons, these relations will be submitted to the same screening as the relations extracted from definitions (see Section 6.).

Two restrictions have been put on the relations we extracted: (*i*) we only collect the pairs made up of words whose lemmas only contain lower case letters. We therefore excluded all proper names, reflexive verbs (*se laver* 'have a wash'), compounds, affixes, morphological components (*ostéo-*), numbers, biochemical terms (*acide icosa-8Z,11Z,14Z-triénoïque* 'icosa-8,11,14-trienoic acid'), etc. (*ii*) The acquisition is limited to entries that are nouns, verbs, adjectives and adverbs. In GLAWI, the morphological relations are separately described for each part of speech (PoS) of the entry. The PoS of the entry is therefore known but we do not have any explicit information on the PoS of the words listed in the morphological sections. For the latter, we assume that the lexemes that share the same lemma are morphologically related to each other through conversion or category shift. Therefore, we add one pair for each PoS of the lemmas included in the morphological sections of the entries. The 32,249 morphological sections present in the articles of the 169,112 entries of GLAWI yielded 125,002 candidate pairs of lexemes.

These pairs are merged with the ones yielded by the definitions (as detailed in Section 4.). We then select the valid pairs from this set of candidates as detailed in Sections 6. and 7.

6. Analogical selection of the derivational relations

When two words such as *développer* and *développement* are in the same morphological relation as two others such as *classer* 'order' and *classement* 'ordering', they form a proportional analogy and could be noted *développer*:*développement*=*classer*:*classement*. In this particular case, the analogy is called formal (Lepage, 1998; Lepage, 2004b; Stroppa and Yvon, 2005; Langlais and Yvon, 2008) because it holds between the strings that realize the four lemma. In other words, the formal difference between *développer* and *développement* is exactly the same as the one that exists between *classer* and *classement*. This difference could be described by the following regex substitution that transforms the first string into the second one `/^(.+)er$/^\1ement$/` or by a non oriented pair of patterns `^(.+)er$:^(.+)ement$` where the variable parts `(.+)` represent the identical substrings in the two patterns. Note that these patterns are not minimal in that they take into account that the first two strings represent infinitive forms of French first conjugation class.

Lepage (2004a; Lepage (2004b) and Stroppa and Yvon (2005) propose two methods that could be used to check whether four strings form an analogy. The latter use a finite state solver to compute the solutions of analogical equations X:Y=Z:?. This method is more general and more complex than the former which can only be used when all four strings are known. However, being in this case, we opted for the former (Lepage, 1998; Lepage, 2004b) since we need to check the existence of analogies between fully specified pairs of strings. This method relies on two observations:

- if A:B=C:D, then the Levenshtein edit distance (Levenshtein, 1966) between A and B is the same as between C and D. In other words, the number of edit operations needed to change A into B is the same as the one needed to change C into D.
- if A:B=C:D, then the characters added or removed from A in order to change it into B are the same as the ones that are added or removed from C to change it into D. In other words, for each character a of the alphabet on which A, B, C and D are built, $|A|_a - |B|_a = |C|_a - |D|_a$ where $|X|_a$ represent the number of occurrences of the character a in the string X.

Moreover, when both conditions are met, the analogy holds with few exceptions (*e.g.*, mirroring or reduplication as in *stressed*:*desserts*=*reward*:*drawer* (Lepage, 2004b)). This method is a very powerful and efficient way to discover analogies because we can associate with each pair of words a signature made up of their edit distance and a description of the differences between the numbers of characters they contain. If two pairs happen to share the same signature, they are likely to form an analogy. And if by chance they do not, they will be taken away by a second phase of selection described in Section 6.. More formally, the signature of a pair (A, B) is:[1]

$$\sigma(A,B) = (d(A,B), |A|_{a_1} - |B|_{a_1}, \ldots, |A|_{a_n} - |B|_{a_n})$$

[1]These signatures can be computed efficiently by means of the `fast_distance` module developed by Yves Lepage based on the algorithm of Allison and Dix (1986) and of the `collections` module.

where $d(A, B)$ is the Levenshtein distance between A and B and $\{a_1, \ldots, a_n\}$ is the alphabet of the language.

We computed the signatures of all the pairs extracted from the GLAWI morphological sections and yielded by the definitions. We only kept the pairs having signatures occurring at least 5 times. This threshold has been chosen because we observed it removes most of the erroneous pairs. However, we did not estimate its recall and precision. This first filter retained 170,370 pairs out of the 2,353,959 initial ones.

The analogical selection is based on the hypothesis that derivational relations form series, namely sets of pairs of words connected by the same derivational relation. For instance, the nominals derived from verbs by suffixation in *-ment* form a derivational series. Derivational series usually contain pairs with formal variations (*i.e.* allomorphy) that are difficult to catch without a more complete semantic analysis. For instance, the *-ment* series contains pairs that instantiate various patterns:

`^(.+)er$:^(.+)ement$` *développer*:*développement*
`^(.+)ir$:^(.+)issement$` *investir*:*investissement* 'invest:investment'
`^(.+)eler$:^(.+)ellement$` *ruisseler*:*ruissellement* 'flow_V:flow_N'
`^(.+)re$:^(.+)ement$` *rendre*:*rendement* 'return_V:yield_N'

For this reason, we use the formal subseries as approximations of the derivational series. In other words, the previous hypothesis is reframed as: derivational relations form formal series, that is series that do not contain formal variations. Fam and Lepage (2018) make use of the same hypothesis to create analogical grids from text corpora. The hypothesis is stated as in (1).

$$\begin{array}{c} P_1^1 : P_1^2 :: \cdots :: P_1^n \\ P_2^1 : P_2^2 :: \cdots :: P_2^n \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ P_m^1 : P_m^2 :: \cdots :: P_m^n \end{array} \iff \begin{array}{l} \forall (i,k) \in \{1,\ldots,n\}^2, \\ \forall (j,l) \in \{1,\ldots,m\}^2, \\ P_i^j : P_i^l = P_k^j : P_k^l \end{array} \quad (1)$$

In our case, we focus only on series of derivational relations. In other words, $n = 2$. Globally, Glawinette could be seen as an analogical grid with at least three differences: we only consider derivational relations; Glawinette identifies the linguistic exponents involved in the derivational relations it contains; Glawinette includes what we may call relations of motivation that connect derivatives to their (semantic) base lexemes obtained from the definitions and from relations between lexemes that belong to the same derivational family.

7. Obtaining more natural exponents from word series

In this work, we pushed the signature-based selection one step further by considering that derivational series must be formally homogeneous, that is made up of pairs of lexemes that share formal properties that correspond to the derivational relation that connects them. For instance, the pairs in Table 2 describe a regular relation between agent nouns in *-eur* and action nouns in *-age*. We observe that all the words in the first column end in *-eur* and conform to a pattern `^(.+)eur$` and all the words in the second column end in *-age* and conform to the pattern `^(.+)age$`. The aim of the pattern-based selection is to find morphologically plausible patterns and select the pairs of lexemes that instantiate them. For instance, the signature of the pairs in Table 2 happens to be the same as the one of the pairs in Table 3 because they all have an edit distance of 4 and differ by the same characters ({`r`, `u`} are deleted and {`a`, `g`} added). For this reason, the pairs in Tables 2 and 3 end up in the same set of relations after signature-based selection even if they do not form an analogy (e.g. *allumeur*:*allumage*≠*doublure*:*doublage*).

`^(.+)eur$`	`^(.+)age$`
allumeur 'igniter'	allumage 'ignition'
atterrisseur ' lander'	atterrissage 'landing'
balayeur 'sweeper'	balayage 'sweeping'
carreleur 'tiler'	carrelage 'tiling'
épandeur 'spreader'	épandage 'spreading'

Table 2: Excerpt of the *-eur*/*-age* series

`^(.+)ure$`	`^(.+)age$`
doublure 'lining'	doublage 'doubling'
boursouflure 'swelling'	boursouflage 'swelling'
rayure 'scratch'	rayage 'scratching'
épluchure 'peeling'	épluchage 'peeling'

Table 3: Excerpt of the *-ure*/*-age* series

The other aim of this pattern-based selection is to identify linguistically plausible exponents. For instance, the minimal patterns for a pair of words such as *productif*:*productivité* is `^(.+)f$:^(.+)vité$` while the linguistic one would be something like X:X*ité*. The pattern-based selection can be seen as a trade-off between the string level and the linguistic abstraction. For the previous example, it provides patterns that include the *-if*:*-ive* allomorphy, namely `^(.+)if$:^(.+)ivité$` which is clearly more natural than `^(.+)f$:^(.+)vité$`.

Series of words display very large numbers of regularities. For instance, in the first column of Table 2, all words end in `r`, `ur` and `eur` but some also start with an `a`, etc. These regularities could be represented by patterns as in the left part of Table 4. We also give their coverage because some are partial. The right part of Table 4 lists similar patterns for the words in *-age* in the third column of Table 2. The number of regularities present in the few pairs in Table 2 is actually very high and many patterns could be added to the tables in Table 4. For instance, we could add patterns to account for the fact that *atterrisseur* and *carreleur* (resp. *atterrissage* and *carrelage*) contain a `rr` substring, or that *allumeur*, *balayeur* and *carreleur* (resp. *allumage*, *balayage* and *carrelage*) contain an `l`, etc.

However, very few of these patterns are relevant linguistically, the most relevant ones being `^(.+)eur$` for the first series of words and `^(.+)age$` for the second. The selection of these relevant patterns relies on the observations that they are in correspondence and that the unspecified sequences `(.+)` represent identical substrings in the

Pattern *-eur*	Cov
^(.+)$	1.0
^(.+)u$	1.0
^(.+)u(.+)$	1.0
^(.+)ur$	1.0
^(.+)e(.+)$	1.0
^(.+)eu(.+)$	1.0
^(.+)eur$	1.0
^a(.+)$	0.4
^a(.+)r$	0.4
^a(.+)u(.+)$	0.4
^a(.+)ur$	0.4
^a(.+)e(.+)$	0.4
^a(.+)eu(.+)$	0.4
^a(.+)eur$	0.4

Pattern *-age*	Cov
^(.+)$	1.0
^(.+)e$	1.0
^(.+)g(.+)$	1.0
^(.+)ge$	1.0
^(.+)a(.+)$	1.0
^(.+)ag(.+)$	1.0
^(.+)age$	1.0
^a(.+)$	0.4
^a(.+)e$	0.4
^a(.+)g(.+)$	0.4
^a(.+)ge$	0.4
^a(.+)a(.+)$	0.4
^a(.+)ag(.+)$	0.4
^a(.+)age$	0.4

Table 4: Excerpt of the formal regularities displayed by the words in *-eur* and *-age* of Table 2 with their coverage. Some regularities are partial and only concern a subset of the words.

pairs of words. For instance, (*i*) the pattern `^a(.+)eur$` in the first series of words corresponds to `^a(.+)age$` in the second one because they match the words of the same pairs in Table 2, namely *allumeur*:*allumage* and *atterrisseur*:*atterrissage* and (*ii*) `(.+)` in the two patterns represents `llum` in the first pair and `tterriss` in the second. Because *allumeur*:*allumage*=*atterrisseur*:*atterrissage* is an analogy, these substrings are also precisely the ones that vary in the *allumeur*:*atterrisseur* and *allumage*:*atterrissage* pairs, that is in the series. This corollary is actually very interesting because it gives us more relevant patterns because it identifies the longest stable substrings in the words of the series. In other words, the difference between *allumeur* and *atterrisseur* is that `llum` in the first word is replaced by `tterriss` in the second one while the initial `a` and the final `eur` remain unchanged which yields the pattern `^a(.+)eur$` very simply by replacing the varying substrings by `(.+)` in these words. A more formal description of this procedure is given in Algorithm 1 where the DIFF function provides a description of the differences that exist between the two strings L_1 and L_2. This difference can be computed by means of the `SequenceMatcher` method of the `difflib` Python module (for a similar use of this method, see (Bernhard, 2010)).

Algorithm 1 Formal pattern of a pair of lexemes

function PATTERN(L_1, L_2)
  $D = \text{DIFF}(L_1, L_2)$
  $P \leftarrow$ ^ ▷ start of string
  for all $O \in D$ **do**
    $O = (\text{Type}, \text{Substring}L_1, \text{Substring}L_2)$
    if Type is `Equal` **then**
      ▷ SubstringL_1 = SubstringL_2
      $P \leftarrow P \oplus \text{Substring}L_1$
    else
      $P \leftarrow P \oplus$ `(.+)`
    end if
  end for
  return $P \oplus$ $ ▷ end of string
end function

Step 1. The patterns that describe all the regularities in a series of words are collected by applying the function PATTERN to all the pairs made up of words of the series, as described in Algorithm 2. In this algorithm, we introduced three conditions to discard the less interesting patterns, namely those that apply to less than MinSize words, those with a coverage lower than MinCover and those that contain a number of variable sequences `(.+)` lower that MinVar or higher that MaxVar. The values we have used for these parameters are MinSize $= 5$, MinCover $= 0.1$, MinVar $= 1$, MaxVar $= 1$.

Algorithm 2 Formal patterns of a series of lexemes

$\mathcal{L} = \{L_1, \ldots, L_n\}$ is a series of wordforms
function LEXEMESERIESPATTERNS($\mathcal{L}$, MinSize, MinCover, MinVar, MaxVar)
  $\mathcal{R} \leftarrow \emptyset$
  for all $i, j \mid 1 \leq i < j \leq n$ **do**
    $P \leftarrow \text{PATTERN}(L_i, L_j)$
    $M \leftarrow \{L \in \mathcal{L} \mid \text{match}(P, L)\}$
    if $|M| \geq$ MinSize **and**
      $\frac{|M|}{|\mathcal{L}|} \geq$ MinCover **and**
      MinVar $\leq$ count(`(.+)`, P) $\leq$ MaxVar **then**
      $\mathcal{R} \leftarrow \mathcal{R} \cup \{P\}$
    end if
  end for
  return $\mathcal{R}$
end function

The patterns that describe all the regularities in a series of words must then be paired with the patterns of the other series of words. For instance, `^a(.+)eur$` must be paired with `^a(.+)age$` in order to form the relevant pattern for the first two pairs of Table 2. As we saw, the `(.+)` in the patterns represent identical substrings these pair of words. We can therefore use them to transform *allumeur* into *allumage* and *atterrisseur* and *atterrissage* by means of a regex substitution `/^a(.+)eur$/a\1age/` where `\1` represents the first memorized substring `(.+)` in the first word. Moreover, when applied to the all the words in the first column of Table 2, this substitution selects the pairs whose words are exactly the ones that match the patterns in correspondence. This pairing procedure is described more formally in Algorithm 3.

Step 2. This first step of pattern selection identifies regularities at the level of the series of words and series of relations. In step 1, all pairs of words that are not part of a selected series of relations are removed. Pattern selection also highlights that most pairs of lexemes are associated with multiple pattern pairs. Table 5 presents the patterns of *verbaliser*:*verbalisation*, *proverbial*:*proverbialement*, *féministe*:*féminisme*. The patterns of *verbaliser*:*verbalisation* are both linguistically relevant as the first indicates that *verbalisation* is coined on the *-at* ending stem of *verbaliser* and the second that *verbalisation* results from a derivation in *-iser* followed by a derivation in *-ion*. On the other hand, only two of the four patterns of *proverbial*:*proverbialement*

		patterns	selected
verbaliser=V 'verbalize'	verbalisation=N 'verbalization'	^(.+)er$:^(.+)ation$ ^(.+)iser$:^(.+)isation$	←
proverbial=A 'proverbial'	proverbialement=R 'proverbially'	^(.+)ial$:^(.+)ialement$ ^(.+)al$:^(.+)alement$ ^(.+)l$:^(.+)lement$ ^(.+)$:^(.+)ement$	←
féministe=A 'feminist'	féminisme=N 'feminism'	^(.+)niste$:^(.+)nisme$ ^(.+)iste$:^(.+)isme$ ^(.+)ste$:^(.+)sme$	←
sarkozysme=N 'sarkosysm'	sarkozyste=N 'sarkosyst'	^(.+)ste$:^(.+)sme$	←

Table 5: Multiple patterns

Algorithm 3 Formal patterns of a series of relations

$\mathcal{S} = \{(L_1, K_1), \ldots, (L_n, K_n)\}$ is a series of relations
$L_1, \ldots, L_n, K_1, \ldots, K_n$ are wordforms
Require: $\forall i, j \in \{1, \ldots, n\}^2, \sigma(L_i, K_i) = \sigma(L_j, K_j)$
function RELATIONSERIESPATTERNS($\mathcal{S}$)
  $\mathcal{P} \leftarrow$ LEXEMESERIESPATTERNS($\{L_1, \ldots, L_n\}$)
  $\mathcal{Q} \leftarrow$ LEXEMESERIESPATTERNS($\{K_1, \ldots, K_n\}$)
  $\mathcal{R} \leftarrow \emptyset$
  for all $P \in \mathcal{P}$ **do**
    for all $Q \in \mathcal{Q}$ **do**
      if $\sigma(P, Q) = \sigma(L_1, K_1)$ **then**
        $\mathcal{R} \leftarrow \mathcal{R} \cup \{(P, Q)\}$
      end if
    end for
  end for
  return $\mathcal{R}$
end function

are relevant:
^(.+)$:^(.+)ement$
^(.+)ial$: ^(.+)ialement$
The second pattern is interesting because it describes that *proverbial* involves a *-ial* variant of the *-al* exponent. For *féministe*:*féminisme*, two out of three patterns are linguistically irrelevant (the only relevant one being ^(.+)iste$:^(.+)isme$) and none is relevant for *sarkozysme*:*sarkozyste*. The method we present is unable to catch the relevant variant ^(.+)yste$:^(.+)ysme$ because it is not frequent enough.

Pattern selection involves a second step that aims to identify the most relevant pattern of a pair of lexemes with respect to its paradigmatic connections. The selected patterns for the pairs in Table 5 are signaled by arrows. As we can see, the selected patterns belong to the linguistically relevant ones, when a relevant one exists. The selection of the relevant relation patterns (P, Q) among the ones yielded in step 1 is based on the assumption that they are made of pairs of relevant lexeme patterns. In this first version of Glawinette, we did not perform a global optimization on the entire set of patterns yielded in the first step. Relevant patterns were selected locally by maximizing the number of pairs of lexemes that fit in the two lexeme patterns they are made of. More precisely, let $\mathcal{R}$ be the set of relation patterns yielded in step 1 for a series of lexemes. In this second step, we select the pattern $(P, Q) \in \mathcal{R}$ that maximizes $|P| + |Q|$, where $|X|$ is the number of lexemes associated with the lexeme pattern X.

A first assessment of the quality of the selected patterns has been performed by two judges who checked 200 pairs of lexemes selected randomly. All the pairs that we annotated were correct, that is made up of morphologically related lexemes. 169 patterns out of 200 (84%) have been considered as linguistically "natural", that is made up of exponents that a morphologist could use to describe the relation between the pair of lexemes. The patterns that were not considered to be natural mainly correspond to verb prefixation (*re-*, *auto-*, *entre-*, etc.) such as ^re(.+)er$:^(.+)er$ (selected pattern for *rehacker*:*hacker* 'rehack:hack') where the infinitive ending er should not have been included. We also found errors in the patterns of derivatives of second conjugation verbs like *réagissable*:*réagir* 'reactable:react'. The selected pattern is ^(.+)ssable$:^(.+)r$ instead of ^(.+)issable$:^(.+)ir$. Only one pair (*parodontiste*:*parodontologiste* 'periodontist:periodontologist') out of 200 had a totally incorrect pattern because it presents an uncommon variation. The annotation also highlighted that these less "natural" patterns could be corrected easily because the errors they present are systematic and because the correction can be performed at the level of the series.

8. Conclusion

We presented in this paper Glawinette, a new French derivational resource created without any manual annotation from the GLAWI machine readable dictionary. In this way, the same processing chain could be used for the Italian and English versions of GLAWI, namely GLAW-IT (Calderone et al., 2016) and ENGLAWI (Sajous et al., 2020). Glawinette will soon be made available under the same license as GLAWI and Wiktionnaire. With 97,293 lexemes and 47,712 relations, the integration of Glawinette in Démonette will significantly increase the number of its entries. Moreover, Glawinette entries are associated with patterns which most often are linguistically valid. This feature will make their manual checking easy because their description is close to the linguistic intuition of the human annotators.

Finally, the derivational families and series described in Glawinette are the basic building blocks for the definition of true derivational paradigms.

9. Acknowledgment

This work benefited from the support of the project DEMONEXT ANR-17-CE23-0005 of the French National Research Agency (ANR). Computations performed to produce Glawinette have been carried out using the OSIRIM platform, administered by IRIT and supported by CNRS, the Région Occitanie, the French Government and ERDF.

10. Bibliographical References

Allison, L. and Dix, T. I. (1986). A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, 23(5):305–310.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, Philadelphia, PA.

Bauer, L. (1997). Derivational paradigms. In *Yearbook of Morphology 1996*, pages 243–256. Springer.

Bernhard, D., Cartoni, B., and Tribout, D. (2011). A task-based evaluation of French morphological resources and tools. *Linguistic Issues in Language Technology*, 5(2).

Bernhard, D. (2010). Apprentissage non supervisé de familles morphologiques: Comparaison de méthodes et aspects multilingues. *Traitement Automatique des Langues*, 51(2):11–39.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bonami, O. and Strnadová, J. (2019). Paradigm structure and predictability in derivational morphology. *Morphology*, 29(2):167–197.

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

Cotterell, R. and Schütze, H. (2018). Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6:33–48.

Creutz, M. and Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Helsinki University of Technology.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fam, R. and Lepage, Y. (2018). Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1060–1066, Miyazaki, Japan.

Goldsmith, J. (2001). Unsupervised learning of the morphology of natural language. *Computational Linguistics*, 27(2):153–198.

Habash, N. and Dorr, B. (2003). A categorial variation database for English. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (NAACL/HLT 2003)*, pages 96–102, Edmonton. ACL.

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

Hathout, N. and Namer, F. (2014b). 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)*, pages 208–219, Marseille. ATALA.

Hathout, N. and Namer, F. (2016). Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Hathout, N. and Namer, F. (2018a). Defining paradigms in word formation: concepts, data and experiments. *Lingue e Linguaggio*, 17(2):151–154.

Hathout, N. and Namer, F. (2018b). La parasynthèse à travers les modèles : des RCL au ParaDis. In Olivier Bonami, et al., editors, *The lexeme in descriptive and theorical morphology*, pages 365–399. Langage Sciences Press.

Hathout, N. and Namer, F. (2019). Paradigms in word formation: what are we up to? *Morphology*, 29(2):153–165.

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

Hathout, N., Sajous, F., and Calderone, B. (2014). 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.

Hathout, N. (2002). From WordNet to CELEX: Acquiring morphological links from dictionaries of synonyms. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1478–1484, Las Palmas de Gran Canaria. ELRA.

Hathout, N. (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. ACL.

Hathout, N. (2009a). Acquisition of morphological families and derivational series from a machine readable dictionary. In Fabio Montermini, et al., editors, *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*, Somerville, MA. Cascadilla Proceedings Project.

Hathout, N. (2009b). *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Habilitation à diriger des recherches, Université de Toulouse 2 - Le Mirail, Toulouse.

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

Hathout, N. (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.

Hathout, N. (2014). Phonotactics in morphological similarity metrics. *Language Sciences*, 46:71–83.

Kyjánek, L. (2018). Morphological resources of derivational word-formation relations. Technical Report 61, ÚFAL, Charles University, Prague.

Langlais, P. and Yvon, F. (2008). Scaling up analogical learning. In *Proceedings of the 22nd international conference on Computational Linguistics (COLING 2008)*, pages 51–54, Manchester, UK.

Lepage, Y. (1998). Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*, volume 2, pages 728–735, Montréal.

Lepage, Y. (2004a). Analogy and formal languages. *Electronic notes in theoretical computer science*, 53:180–191.

Lepage, Y. (2004b). Lower and higher estimates of the number of true analogies between sentences contained in a large multilingual corpus. In *Proceedings of the 20th international conference on Computational Linguistics (COLING-2004)*, pages 736–742, Genève.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10(8):707–710.

Martin, R. (1992). *Pour une logique du sens*. Linguistique nouvelle. Presses universitaires de France, Paris.

Namer, F., Barque, L., Bonami, O., Haas, P., Hathout, N., and Tribout, D. (2019). Demonette2 - une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26e conférence annuelle sur le traitement automatique des langues naturelles (TALN-2003)*, pages 233–243, Toulouse.

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues : L'analyseur DériF*. Hermès Science-Lavoisier, Paris.

Namer, F. (2013). A rule-based morphosemantic analyzer for French for a fine-grained semantic annotation of texts. In Cerstin Mahlow et al., editors, *SFCM 2013*, CCIS 380, pages 93–115. Springer, Heidelberg.

Plénat, M. (2005). Brèves remarques sur les déverbaux en *-ette*. In Frédéric Lambert et al., editors, *La syntaxe au coœur de la grammaire. Recueil offert en hommage pour le 60e anniversaire de Claude Muller*, pages 245–258. Presses universitaires de Rennes, Rennes.

Roché, M., Boyé, G., Hathout, N., Lignon, S., and Plénat, M. (2011). *Des unités morphologiques au lexique*. Hermès Science-Lavoisier, Paris.

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, England.

Sajous, F., Calderone, B., and Hathout, N. (2020). ENGLAWI: From Human- to Machine-Readable Wiktionary. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France.

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

Van Marle, J. (1985). *On the Paradigmatic Dimension of Morphological Creativity*. Foris, Dordrecht.

Vidra, J., Žabokrtskỳ, Z., Ševčíková, M., and Kyjánek, L. (2019). Derinet 2.0: Towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, pages 81–89, Prague.

Zeller, B. D., Snajder, J., and Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for german. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1201–1211, Sofia, Bulgaria.