Glawinette: a linguistically motivated derivational description of French acquired from GLAWI
Nabil Hathout, Franck Sajous, Basilio Calderone, Fiammetta Namer

To cite this version:

HAL Id: halshs-02929815
https://halshs.archives-ouvertes.fr/halshs-02929815
Submitted on 9 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Glawinette: a linguistically motivated derivational description of French acquired from GLAWI

Nabil Hathout¹, Franck Sajous¹, Basilio Calderone¹, Fiammetta Namer²
¹CLLE, CNRS & Université de Toulouse Jean Jaurès, ²ATILF, 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 Wiktomicnaire, 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 pr
ince ‘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’Adv, the latter being the relational adjective of prince and other are indirect like introducer: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.

Figure 1: The derivational families of prince ‘prince’ and introduire ‘introduce’ in Glawinette

The second projection Glawinette provides is the deriva-
tional 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 and 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 ^(.+)$=N:^(.+)$ion$=N where (+) represents the initial sequence shared by the pair of words. Table 1 presents an excerpt of the series of this pair.

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

<table>
<thead>
<tr>
<th>^(.+)eur$=N</th>
<th>^(.+)ion$=N</th>
</tr>
</thead>
<tbody>
<tr>
<td>acteur</td>
<td>action</td>
</tr>
<tr>
<td>animateur</td>
<td>animation</td>
</tr>
<tr>
<td>classificateur</td>
<td>classification</td>
</tr>
<tr>
<td>colonisateur</td>
<td>colonisation</td>
</tr>
<tr>
<td>directeur</td>
<td>direction</td>
</tr>
<tr>
<td>decentralisateur</td>
<td>decentralisation</td>
</tr>
<tr>
<td>dépresser</td>
<td>dépression</td>
</tr>
<tr>
<td>éditeur</td>
<td>édition</td>
</tr>
<tr>
<td>expositeur</td>
<td>exposition</td>
</tr>
<tr>
<td>formateur</td>
<td>formation</td>
</tr>
<tr>
<td>réacteur</td>
<td>réaction</td>
</tr>
<tr>
<td>réviser</td>
<td>révision</td>
</tr>
</tbody>
</table>

Figure 2: Series are characterized by 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 ‘form’). 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$. Glawinette contains 5400 series of relations, the largest being the noun-verb conversion (i.e., zero-derivation) ^(.+)$=N:^(.+)$er$=V with 3918 pairs. The smallest series contain pairs with idiosyncratic written forms such as naïf:naïveMasc naiveFem.

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 summarized 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 Laus, 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 years, French lacked a large-scale derivational resource 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, déveleopppe ‘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 occurs 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 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 icosaoctadecaénoïque ‘icoso-octadecenoic 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 développeur:développement=classer:classement. In this particular case, the analogy is called formal (Lepage, 1998, Lepage, 2004a, Lepage, 2004b, Stroppa and Yvon, 2005, Langlas 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$:^\1ement$ 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, 1965) 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:dessert=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

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

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$_1$, ..., 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:

\[
\begin{align*}
&\overset{(.+)}{\text{er}}:\overset{(.+)}{\text{ement}}\overset{\text{dév}}{\text{évelopper}} \overset{\text{développement}}{\text{développement}}, \\
&\overset{(.+)}{\text{ir}}:\overset{(.+)}{\text{issement}} \overset{\text{investir}}{\text{investissement}}, \\
&\overset{(.+)}{\text{eler}}:\overset{(.+)}{\text{ellement}} \overset{\text{ruisseler}}{\text{ruissellement}}, \\
&\overset{(.+)}{\text{re}}:\overset{(.+)}{\text{ement}} \overset{\text{rendre}}{\text{rendement}}, \\
&\overset{(.+)}{\text{er}}:\overset{(.+)}{\text{ement}} \overset{\text{allumenter}}{\text{allumage}}.
\end{align*}
\]

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).

\[
P_1^1 : P_2^1 : \ldots : P_n^1 \\
P_2^1 : P_2^2 : \ldots : P_n^2 \\
\vdots : \vdots : \vdots \\
P_m^1 : P_m^2 : \ldots : P_m^n \\
\Leftrightarrow \forall (i, k) \in \{1, \ldots, n\}^2, \\
\forall (j, l) \in \{1, \ldots, m\}^2,
\]

In this 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 morpho-

<table>
<thead>
<tr>
<th>^(.+)eur$</th>
<th>^(.+)age$</th>
</tr>
</thead>
<tbody>
<tr>
<td>allumeur 'igniter'</td>
<td>allumage 'ignition'</td>
</tr>
<tr>
<td>atterrisseur 'lander'</td>
<td>atterrissage 'landing'</td>
</tr>
<tr>
<td>balayeur 'sweeper'</td>
<td>balayage 'sweeping'</td>
</tr>
<tr>
<td>carreleur 'tiler'</td>
<td>carrelage 'tiling'</td>
</tr>
<tr>
<td>épandeur 'spreader'</td>
<td>épandage 'spreading'</td>
</tr>
</tbody>
</table>

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

Logically 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. allumage:allumage\#double:doubleage).

<table>
<thead>
<tr>
<th>^(.+)eur$</th>
<th>^(.+)age$</th>
</tr>
</thead>
<tbody>
<tr>
<td>doublure 'lining'</td>
<td>doublage 'doubling'</td>
</tr>
<tr>
<td>boursouflure 'swelling'</td>
<td>bourselage 'swelling'</td>
</tr>
<tr>
<td>rayure 'scratch'</td>
<td>rayage 'scratching'</td>
</tr>
<tr>
<td>épiluchure 'peeling'</td>
<td>épiluchage 'peeling'</td>
</tr>
</tbody>
</table>

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 produit:productivité is ^(.+)if$:^(.+)ivité$ while the linguistic one would be something like X:Xifie. 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 ^(.+)if$:^(.+)ivité$.

Series of words display very large numbers of regularities. For instance, in the first column of Table 2 all words end in rr, 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 2. We also give their coverage because some are partial. The right part of Table 2 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 2. 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
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 same pairs in Table 2, namely allumeur:allumage and atterrisseur:atterrissage and (ii) (.) in the two patterns represents llum in the first pair and tterrain in the second. Because allumeur:allumage=atterrissureen:atterrissage is an analogy, these substrings are also precisely the ones that vary in the allumeur:atterrissieur 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 tterrain in the second 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

```python
function PATTERN(L_1, L_2)
    D = DIFF(L_1, L_2)
    P ← ↦ start of string
    for all O ∈ D do
        O = (Type, Substring L_1, Substring L_2)
        if Type is Equal then
            ↦ Substring L_1 = Substring L_2
            P ← P ⊕ Substring L_1
        else
            P ← P ⊕ (.)
        end if
    end for
    return P ⊕ ↦ end of string
end function
```

**Algorithm 2** Formal patterns of a series of lexemes

$L = \{L_1, \ldots, L_n\}$ is a series of wordforms

```python
function LEXEMESERIESPATTERNS(L, MinSize, MinCover, MinVar, MaxVar)
    R ← ∅
    for all i, j | 1 ≤ i < j ≤ n do
        P ← PATTERN(L_i, L_j)
        M ← \{L ∈ L | match(P, L)\}
        if |M| ≥ MinSize and |M| ≥ MinCover and MinVar ≤ count(., P) ≤ MaxVar then
            R ← R ∪ \{P\}
        end if
    end for
    return 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 into 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 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 2.

**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 1. 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.

**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 3 presents the patterns of verbaiser:verbalisation, proverbial:proverbialment, féministe:féminisme. The patterns of verbaiser:verbalisation 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:proverbialment "...(3882 characters)..."
The second pattern is interesting because it describes patterns made up of exponents that a morphologist could use to describe the relation between the pair of lexemes. The selected patterns belong to the linguistically relevant ones, as expressed by the correction can be performed at the level of the series. Moreover, Glawinette entries are associated with patterns that were not considered to be natural mainly because the errors they present are systematic and because 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 for the Italian and English versions of GLAWI, namely GLAW-IT (Calderone et al., 2016) and ENGLISH (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


