# Giving Lexical Resources a Second Life: Démonette, a Multi-Sourced Morpho-Semantic Network for French

Nabil Hathout, Fiammetta Namer

# Giving Lexical Resources a Second Life: Démonette, a Multi-Sourced Morpho-Semantic Network for French

Nabil Hathout*, Fiammetta Namer**
*CLLE, Université de Toulouse, **Université de Lorraine & ATILF
Nabil.Hathout@univ-tlse2.fr, Fiammetta.Namer@univ-lorraine.fr

May 23, 2016

## 1 Introduction

This paper presents Démonette, a derivational morphological network designed for the description of French. Démonette features an original architecture that enables its use as a formal framework for the description of morphological analyses and as a repository for existing lexicons. It was fed with a variety of resources, which all were already validated. The harmonization of their content into a unified format offers them a second life. Moreover, they are enriched with new properties provided these can be deduced from their content. Démonette is released under a Creative Commons license. It is usable for theoretical and descriptive research in morphology, as a source of experimental material for psycholinguistics, natural language processing (NLP) and information retrieval (IR), where it fills a gap, since French lacks a large-coverage derivational resources database, similar to CELEX Baayen et al. (1995) or DerivBase Zeller et al. (2013).

In its current state, Démonette consists of information coming from four different sources. They have been added in three successive stages. Overall, Démonette contains 108 888 entries. The entries are a morphological relations, that is pairs of morphologically related words ($W_1$, $W_2$). $W_1$ and $W_2$ belong to the same derivational family and one of them at least is a derived word. Each entry associates a set of structural, morpho-semantic and morpho-phonological descriptions to a morphological relationship. Derived words described in Démonette include deverbal action

nouns (*essorage* 'spin'$_{\text{N}}$), agent nouns (*ramasseur* 'collector') and deverbal adjectives (*productif* 'productive'). Demonette also contains simplex verb predicates (*construire* 'build').

The addition of new entries and the incorporation of new resources generate new information that emerge from the combination of the new data with the descriptions already present in Démonette. In this article, we present the computational and linguistic challenges of the integration of new resources into Démonette and illustrate them with the incorporation of two lexicons: Verbaction and Lexeur.

The remainder of the article is organized as follows: Section 2 introduces Démonette's main features. Section 3 presents the resources used to create the current version of Démonette: DériF, Morphonette, Verbaction and Lexeur. We then discuss some aspects of the integration of the two last into Démonette (Section 4). Finally, in Section 5, we review the adaptability of Démonette with the descriptive requirements of the derivational morphology of French.

## 2 Démonette

Démonette Hathout & Namer (2014) is a general resource designed for the description of word formation (WF) of French. It is eventually intended to partially fill the lack of broad-coverage morphological resources of French as they exist for other languages such as DerivBase for German Zeller et al. (2013) or CELEX Baayen et al. (1995) for English, German and Dutch. Démonette has an original structure since it is a directed graph, where vertices represent lexemes and edges represent derivational relations. In its current version (1.3), it only contains relations between members of the same family.

As illustrated in Figure 1, Démonette includes both derived words (*décorateur* 'decorator'$_{\text{N.masc}}$, *decoratrice* 'decorator'$_{\text{N.fem}}$, *décoration* 'decoration', *décoratif* 'decorative') and simplex words (*décorer*). Simplex words are included only if they are connected to a derived word. Each edge in the graph represents a derivational re-

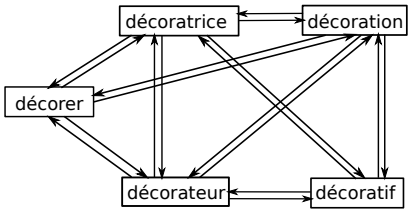


Figure 1: Derivational relations between the members of the *décorer* 'decorate' family

lation described by an entry in the database. These relations are characterized by three properties. The first two are combined in one feature.

We first distinguish direct and indirect relations. **Direct relations** connect a base and its derivatives (*décorer* ↔ *décoration*). One remarkable feature of Démonette is that it also contains **indirect relations** between lexemes of the same derivational family that do not derive one from the other if the relations are semantically predictable: *décoration* and *décorateur* are connected by an indirect relation because decoration is what decorators do. Indirect relations are very useful in families such as {*prédation* 'predation', *prédateur* 'predator'$_{\text{N.masc}}$, *prédatrice* 'predator'$_{\text{N.fem}}$} because there is no verb **préder* 'predate' in French.

Relations can also be characterized by their **orientation**. As we said, Démonette is a directed graph where a relation $W_1 \leftarrow W_2$ describes the **morphological motivation** of $W_1$ with respect to $W_2$, that is the potentiality to construct (and analyze) $W_1$ starting from $W_2$. Most often, if $W_1$ can be motivated with respect to $W_2$, then we can also motivate $W_2$ with respect to $W_1$. In other words, most lexemes are connected to each other in both directions. We end up with three values for the combined-feature **orientation**: direct descending relations connect a derived lexeme to its base or to a more distant ascendant (*décorer* ← *décorateur*); direct ascending relations connect a lexeme to its derivative or to a more distant descendant (*décorateur* ← *décorer*); indirect relations are bi-directed (*décorateur* ↔ *décoratrice*).

The third property of the derivational relations is their **complexity**. Direct relations are simple if they correspond to single derivational operations (*chanteur* 'singer'$_{\text{N.masc}}$ ← *chanter* 'sing'; *chanter* ← *chanteur*). We also consider as simple the indirect relations between words if they are connected through a path of two simple direct relations (*chanteur* ← *chanteuse* 'singer'$_{\text{N.fem}}$) or if they belong to series of words connected by simple relations (*prédateur* ← *prédation*). Exceptional derivational relations are labelled as lexical (see Section 4.5). All other derivational relations are complex. For instance, two derivations are needed to connect *progresser* 'progress'$_{\text{V}}$ to progressivité 'progressivity': *progresser* > *progressif* 'progressive' > *progressivité*. Table 1 summaries the features used to characterize derivational relations in Démonette.

Démonette also provides a wide range of information about the relations and the lexemes it contain. Derivational relations being identified by the words they connect ($W_1 \leftarrow W_2$) and by their properties, Démonette gives the **written form** and the **grammatical category** of their lemma: POS and inflectional features in the EAGLE/GRACE format Rajman et al. (1997). For instance, in the relation *production* 'production' ← *produire* 'produce', the word forms are given the following values of **written form** / **grammatical category** respectively:

production / Ncms produire / Vmn----

In addition to **orientation** and **complexity**, the morphological properties of the relations are characterized by the **type** of the construction: pref(ixation), suf(fixation), conv(ersion); the possible derivational **exponent** of $W_1$ and $W_2$ (i.e., its suffix or prefix value); the derivational written form of $W_1$ and $W_2$'s **root**s, that is the sequence left after the affix (if any) is removed from each $W_i$. In the case of a simple descendant relation between $W_1$ and $W_2$, $W_1$'s **root** value corresponds $W_2$'s stem. This comprehensive set of features allows the detailed description of a variety of derivational relations including regular affixation such as *conceptrice* 'conceptor'$_{\text{N.fem}}$ ← *concevoir* 'conceive' or *formation* 'training' ← *formateur* 'trainer'.

One more original feature of Démonette is the morpho-semantic description of the words and the relations. The lexemes that participate in the relations are assigned to morpho-semantic types. Six types are used in the current version :

- @ for predicates (*décorer*, *aboyer* 'bark'$_{\text{V}}$),
- @ACT for action noun (*décoration*, *aboiement* 'bark'$_{\text{N}}$, *audition* 'hearing'),
- @RES for result noun (*décoration*, *aboiement*, *audition* 'audition'),
- @AGM for agentive masculine noun (*décorateur*, *aboyeur* 'barking (dog)'$_{\text{N.masc}}$, *auditeur* 'listener'$_{\text{N.masc}}$),
- @AGF for agentive feminine noun (*décoratrice*, *aboyeuse* 'barking (dog)'$_{\text{N.fem}}$, auditrice 'listener'$_{\text{N.fem}}$),
- @PRP for property adjectives (*décoratif*, *auditif* 'auditive').

In addition, entries describe the constructed meaning of their first word ($W_1$) by **concrete** and **abstract definitions**. For instance, the constructed meaning of *danseuse* when it is considered in its relationship with *danser* could be defined as in

| **Entry** | **Complexity** | **Orientation** |
|---|---|---|
| chanteur ← chanter | simple | descendant |
| chanter ← chanteur | simple | ascendant |
| chanteur ← chanteuse | simple | indirect |
| prédateur ← prédation | simple | indirect |
| progressivité ← progresser | complex | descendant |
| progressivité ← progression | complex | indirect |

Table 1: Démonette describes different types of derivational relations

| entry | $\mathbf{typ_1}$ | $\mathbf{exp_1}$ | $\mathbf{root_1}$ | $\mathbf{typ_2}$ | $\mathbf{exp_2}$ | $\mathbf{root_2}$ |
|---|---|---|---|---|---|---|
| conceptrice ← concevoir | suf | rice | concept | – | – | – |
| formation ← formateur | suf | ion | format | suf | eur | format |

Table 2: Structural description

| | | |
|---|---|---|
| 1 | con | (agent féminin OR instrument) de danser<br>'(feminine agent OR instrument) of $\text{danse}_{\text{V}}$' |
| 2 | con | celle qui a pour correspondant masculin de danseur<br>'which has $\text{danser}_{\text{N.Masc}}$ as a masculine counterpart' |
| 3 | con | (agent féminin OR instrument) de la danse<br>'(feminine agent OR instrument) of $\text{danse}_{\text{N}}$' |
| 4 | abs | (agent féminin OR instrument) de @<br>'(feminine agent OR instrument) of @' |
| 5 | abs | celle qui a pour correspondant masculin de @AGM<br>'which has @AGM as a masculine counterpart' |
| 6 | abs | (agent féminin OR instrument) de @ACT<br>'(feminine agent OR instrument) of @ACT' |

Table 3: Concrete and abstract definitions

row 1 (concrete definition) and row 3 of Table 3 (abstract definition where the base is replaced by its morpho-semantic type; all feminine agent or instrument nouns share this abstract definition). Démonette is based on a cumulative conceptualizion of the constructed meaning of derived lexemes: each morphological relation contributes to the meanings of the words it connects; the morphological meaning of a word is an aggregation of these redundant elementary meanings. For instance, *danseuse* is also defined as in rows 2 and 3 (concrete definitions with respect to masculine agent noun *danseur* and action noun *danse* 'danse'$_{\text{N}}$) and as in rows 5 and 6 (abstract definitions).

Démonette is an open resource licensed under a Creative Commons license. It is fed by descriptions from various existing derivational bases. However, the information contributed by the different resources is not "dissolved" in the database since Démonette records the origin of all the information it contains as illustrated in Figure 2. The first version of Démonette was created from (*i*) the results of the parsing of the lemmas of the TLFnome lexicon[1] by the morphological analyzer DériF and (*ii*)

[1]TLFnome is a lexicon created from the TLF word list. It contains 97,000 lemmas and is

the Morphonette lexicon. We then added entries from Verbaction and from Lexeur.

| | | | |
|---|---|---|---|
| Form_1 | doseuse | Process_1 | suf |
| Source_Form_1 | tlfnome | Exponent_1 | euse |
| Form_2 | dosage | Source_Constr_1 | demonette |
| Source_Form_1 | tlfnome | Process_2 | suf |
| Cat_1 | Ncfs | Exponent_2 | age |
| Source_Cat_1 | tlfnome | Source_Constr_2 | demonette |
| Cat_2 | Ncms | Type_1 | @AGF |
| Source_Cat_2 | tlfnome | Source_Type_1 | demonette |
| Complexity | simple | Type_1 | @RES |
| Source_Complexity | demonette | Source_Type_1 | demonette |

Figure 2: Excerpt of the *doseuse* 'dosing device'$_{\text{N.fem}}$ ← *dosage* 'dosage' entry. All information is sourced.

Démonette is a highly redundant database. A morphological relation described in more than one resource has as many entries as there are sources that contain it. Moreover, many descriptions can be deduced from other relations because most relations are symmetrical or could be recovered transitively. However, we consider this redundancy as beneficial because it eases the creation of the resource, its update, and its use. Extracting the relations originating from one source is for instance straightforward. Démonette is robust since only 6 fields are required in a entry $W_1 \leftarrow W_2$: the forms, categories and morpho-semantic types of the two words. All the other can remain empty if the information is missing or not relevant. Another interesting feature of Démonette is its flexibility. Its format is open and can be extended with new fields in order to accommodate other types of information.

# 3 Four sources of derivational descriptions

The first version of the network was built from DériF and Morphonette. It contains suffixed action and agent nouns, formed by *-age*, *-ment*, *-ion*, *-eur*, *-euse* and *-rice*, *-if* suffixed adjectives denoting properties, as well as the corresponding verb base, when available.

Démonette's current version[2] contains two additional resources: Verbaction and

extremely high in quality by virtue of many manual reviews. The XML version of this lexicon, called Morphalou, is available from the ATILF-CNRS laboratory at www.cnrtl.fr/lexiques/morphalou/.

[2]Démonette is available on two repositories:
http://redac.univ-tlse2.fr/lexiques/demonette.html
https://www.ortolang.fr/market/lexicons/demonette

Lexeur. Verbaction is a lexicon of action nouns; Lexeur's entries connects an agent noun derived by *-eur* suffixation with its morphological base and corresponding action nouns. Both resources were designed for NLP and IR. Although they contain information compatible with those present in the previous version of Démonette, they each have some original properties and a specific structure that requires the development of a dedicated program of conversion and completion. While ensuring that all the initial information is stored in the Démonette network, the program also calculates all the information needed to fill in all the fields of Démonette.

## 3.1 DériF

DériF Namer (2009, 2013) is a morphological analyzer that implements WF rules developed by linguists. One of its major features is that its analyses are controlled by a set of exceptions that account for some of the irregularities that have accumulated during the evolution of the lexicon. Another remarkable characteristics is that DériF provides each derived word with a (concrete) definition, that is, a phrase that expresses its morphologically constructed meaning with respect to its base when the word is formed through derivation, as in (1a), or with respect to its bases in the case of compounding. These definitions are reminiscent of those of WordNet Miller et al. (1990); Fellbaum (1999).

DériF analyzes POS-tagged lemmas. It recursively applies the WF rules until a non-decomposable unit is identified, i.e., a string in which no affix or compounding element can be found and whose part of speech makes it unlikely to be a converted word. DériF provides a list of the morphological antecedents of the analyzed word and a morphological definition, as in (1b). The first element in the list is the analyzed word. In (1a), logical "OR" indicates an ambiguous meaning. DériF also provides a set of features that reflect the constraints imposed by the morphological rules Namer (2002); Namer et al. (2009).

(1) a. *enneigement/NOM* ← *enneiger/VER (action OR résultat de l'action) de enneiger* '(action OR result of the action) of covering with snow'

b. *(enneigement/NOM, enneiger/VER, neige/NOM)* '(snow cover, cover with snow, snow)'

## 3.2 Morphonette

Morphonette is a French lexical network based on a relational and paradigmatic conceptualizion of morphology Hathout (2011). The morphological properties of lexemes

are described by their derivational family and series. For example, a derivative such as *modifiable* 'modifiable' belongs to its derivational family, which encompasses the lexemes in (2a), and to its derivational series, which contains all *-able* derivatives (2b).

(2) a. *modifier*$_{\mathrm{V}}$, *modification*$_{\mathrm{N}}$, *modificateur*$_{\mathrm{N}}$, *modificatif*$_{\mathrm{A}}$, *modifiant*$_{\mathrm{A}}$... 'modify, modification, modifier, amending, modifying'

b. *amplifiable*$_{\mathrm{A}}$, *glorifiable*$_{\mathrm{A}}$, *identifiable*$_{\mathrm{A}}$, *définissable*$_{\mathrm{A}}$, *différenciable*$_{\mathrm{A}}$... 'amplifiable, glorifiable, identifiable, definable, differentiable'

(3) *(modifiable*$_A$*, modificateur*$_N$*,*
*{amplifiable*$_A$*, glorifiable*$_A$*, identifiable*$_A$*...})*

Morphonette was constructed from TLFnome. It is composed of filaments such as in (3), i.e. triplets $(w, r, s_r(w))$, where $w$ is the entry, $r$ is a member of the derivational family of $w$ and $s_r(w)$ is the derivational series of $w$ with respect to $r$. Filaments are interesting because they describe the relations between a derivative and (*i*) its base (direct relations), (*ii*) the members of its derivational family (indirect relations) and (*iii*) the members of its derivational series. The objective of Morphonette is to discover and represent the derivational relations which exist between the lemmas of TLFnome. These relations are searched in neighborhoods defined by the Proxinette measure Hathout (2009, 2014). Morphonette contains 29 310 words and 96 107 filaments.

## 3.3 Verbaction

Verbaction Hathout & Tanguy (2002) is one of the first freely distributed derivational lexicons for French. Intended for NLP, this lexicon contains 9 393 noun-verb pairs such that (*i*) the noun is morphologically related to the verb in synchrony or historically and (*ii*) the noun can be used to express the action denoted by the verb. This resource can therefore be used to identify nominal and verbal expression of variants of the same information (4). Verbaction has been fully manually checked.

(4) a. Les ingénieurs **développent** des logiciels de veille économique. 'The engineers develop of business intelligence software.'

b. Le **développement** de logiciels constitue la principale activité de la société. 'Software development is the main activity of the company.'

Verbaction involves a great variety in derivational processes including a large number of suffixes (*-ade*, *-age*, *-aison*, *-ance*, *-ée*, *-ence*, *-ette*, *-ie*, *-ment*, *-ion*, *-ure*, etc.), various types of conversion, and a great heterogeneity as far as morphological orientation is concerned, cf. section 4.5 On the other hand, Verbaction is very consistent on the semantic level since all nouns can denote action and are related to their corresponding verbs in the same way.

## 3.4 Lexeur

**General description.** Lexeur is a derivational lexicon of *-eur* suffixed agent nouns, consisting of 4 188 entries. Lexeur was initially designed study argument-structure similarities between *-eur* masculine agent nouns and their morphologically related action-denoting predicate Fabre et al. (2004). These predicates are either verbs (*danser* for *danseur*), or nouns (*football* for *footballeur* 'footballer'). Lexeur entries are triplets ($W_1$, $W_2$, $W_3$) describing the relation between the agent noun ($W_1$, e.g. *danseur*), the corresponding predicate ($W_2$, e.g. *danser*) and the associated action noun ($W_3$, e.g. *danse*) Hathout & Fabre (2002). In some entries, $W_2$ or $W_3$ are missing: for instance, *délateur* 'informer' has no base verb nor base noun; Lexeur encode no action noun for *confiseur* 'confectioner'.

**Feminine agent nouns in *-euse* and *-rice*.** Feminine agent nouns have been added to each Lexeur entry as part of Lexeur-to-Démonette migration. The program we designed predicts all the corresponding feminine agent noun that belong to the family of the masculine agent noun, its base and its related action nouns. The goal is to complement the derivational paradigm with possible, morphologically well-formed feminine agent nouns, be they attested in dictionaries or not.

In French, masculine deverbal agent nouns suffixed in *-eur* correspond to feminine agent noun is suffixed either with *-euse*, such as *danseuse* 'dancer'$_{\text{N.fem}}$ (feminine noun corresponding to *danseur*), or with *-rice*, such as *accusatrice* 'accuser'$_{\text{N.fem}}$ (feminine noun corresponding to *accusateur* 'accuser'$_{\text{N.masc}}$). In the beginning, the *-euse*/*-rice*

| | |
|---|---|
| façonnage/Ncms 'shaping' | façonner/Vmn---- 'shape' |
| facturation/Ncfs 'invoicing' | facturer/Vmn---- 'invoice' |
| fauche/Ncfs 'mowing' | faucher/Vmn---- 'reap' |
| fermeture/Ncfs 'locking' | fermer/Vmn---- 'close' |
| ferraillerie/Ncfs 'endless quarell' | ferrailler/Vmn---- 'quarell' |

Figure 3: Verbaction noun-verb entries.

suffix variation had historical grounds: deverbal feminine agent nouns are suffixed in *-rice* if they are inherited from Latin and if their etymon is derived from a verb supine stem in *-tum* (e.g. Latin: *accūso*, *āvi*, *ātum*, *āre* 'accuse'$_{\text{V}}$ > *accūsātrix* 'accuser'$_{\text{N.fem}}$ etymon for French: *accusatrice*$_{\text{N.fem}}$). The other feminine agent nouns borrowed from Latin or (re)constructed in French are suffixed with *-euse*, though a few of them are suffixed with *-esse*, e.g. *vengeresse* 'avenger'$_{\text{N.fem}}$, corresponding to the masculine noun *vengeur*$_{\text{N.masc}}$.

In most cases, our program uses graphical evidences on the *-eur* suffixed nouns to predict the value of the corresponding feminine agent nouns. The feminine nouns ends in *-rice* when the verb stem occurring in masculine noun ends with the one of the graphic sequences *-at* (*accus**at**eur* > *accusatrice*), *-it* (*déb**it**eur* 'debtor'$_{\text{N.masc}}$ > *débitrice* 'debtor'$_{\text{N.fem}}$) or *-ut* (*distrib**ut**eur* 'distributor$_{\text{N.masc}}$' > *distributrice* 'distributor$_{\text{N.fem}}$'), but does not occur in the inflected verb forms, in fact, in its infinitive form (*accuser* 'accuse', *devoir* 'owe', *distribuer* 'distribute'). Besides these main cases, an additional small set of sequences remnant of the Latine supine is used to predict other *-rice* feminine counterparts, when the sequence never occur on the base verb forms: *-ct-*, as with *condu**ct**rice*$_{\text{N.fem}}$ connected to *condu**ct**eur* 'driver$_{\text{N.masc}}$' (from *conduire*,$_{\text{V}}$) or *-pt-*, as with *rece**pt**rice*$_{\text{N.fem}}$, related to *réce**pt**eur* 'receiver$_{\text{N.masc}}$' (from *recevoir*$_{\text{V}}$), and a few unmarked isolated cases: that of *inventrice* 'inventor$_{\text{N.fem}}$'connected to the masculine noun *inventeur*, both derived from *inventer* 'invent$_{\text{V}}$', or that of agent nouns related to verbs derived from *tenir* 'hold'$_{\text{V}}$ (*obtenir* 'obtain'$_{\text{V}}$ > *obtenteur*$_{\text{N.masc}}$ / *obtentrice*$_{\text{N.fem}}$ , *détenir* 'possess'$_{\text{V}}$ > *détenteur*$_{\text{N.masc}}$ / *détentrice*$_{\text{N.fem}}$). The form of the feminine agent noun can also be predicted from the related action noun: the former is suffixed in *-rice* when the latter is suffixed in *-ion*; it is suffixed in *-euse* when the action nouns is suffixed in *-age*, *-ment*, *-ure*, *-erie*, etc. Hence, *obtentrice* and *inventrice* are the feminine agent nouns related to the action noun *obtention* and *invention* in the family of *obtenteur* (resp. *inventeur*). Some masculine agent nouns have two feminine counterparts, one in *-euse* and the other one in *-rice*, the latter form being more frequent (*enquêteuse/enquêtrice* 'investigator'$_{\text{N.fem}}$). All the other masculine agent nouns in *-eur* have a feminine correspondent in *-euse*.

# 4 Enhancing Démonette with Verbaction and Lexeur

Incorporating an NLP resource into a network that records linguistic descriptions requires several adaptations because their structures reflect the diversity of their purposes and do not match perfectly. More precisely, both Verbaction and Lexeur are primarily lexical semantic resources, whereas Démonette was designed as a database able to describe a fragment of French derivational morphology. The design of Ver-

baction and Lexeur follows an onomasiological "meaning first" perspective, where the basic goal is to gather action nouns (resp. agent nouns) morphologically related to verb predicates, no matter which of the verb or the noun is derived from the other, and irrespective of their morphological structures. The development of converters for Verbaction and Lexeur has therefore to solve several kinds of problems resulting from these conceptual divergences. As shown below, this includes the detection of infrequent WF rules such as the *-ing* suffix in *zapping*, the treatment of the back-formation process Becker (1994), as used to form the verb *hydroplaner* 'hydroplane' on the base noun *hydroplanage* 'hydroplaning', cf. Namer (2012), the identification of complex parenthood relationships, e.g. between *bitumisation* 'asphaltisation' and *bituminer* 'asphalt'$_{\mathrm{V}}$, and the detection of undecidable orientation cases, triggered by the conversion process, e.g. between *analyser* 'analyze'$_{\mathrm{V}}$ and *analyse* 'analysis'$_{\mathrm{N}}$.

## 4.1 Sparse data

Seven highly productive suffixes deriving deverbal nouns: *-age*, *-ment*, *-ion*, *-eur*, *-euse*, *-rice*, and deverbal adjectives: *-if*, have already been dealt with in the first version of Démonette. Parsing the nouns in Verbaction and in Lexeur formed by these suffixation is therefore easy. In addition to these suffixes, the migration program has to analyze carefully many particular configurations, triggered by the large amount of low-productive suffixation rules used to derive action nouns in Verbaction, or to connect them to agent nouns in Lexeur. Eighteen new affixes have been found:

> *-ade* (*bousculade* 'rush'), *-aille* (*retrouvaille* 'reunion'), *-aire* (*commentaire* 'comment'), *-aison* (*combinaison* 'combination'), *-ance* (*accoutumance* 'dependecy'), *-ande* (*offrande* 'donation'), *-ange* (*louange* 'praise'), *-ence* (*adhérence* 'adhesion'), *-erie* (*cajolerie* 'cuddle'), *-et* (*ricochet* 'ricochet'), *-ette* (*trempette* 'dipping'), *-eur* (*erreur* 'mistake'), *-ice* (*exercice* 'exercise'), *-ie* (*garantie* 'guarantee'), *-ing* (*kidnapping* 'kidnapping'), *-is* (*arrachis* 'uprooting'), *-ise* (*chapardise* 'pilfering'), *-isme* (*exorcisme* 'exorcism'), *-ité* (*mendicité* 'begging'), *-ment* (*miaulement* 'mewing'), *-oire* (*interrogatoire* 'questioning'), *-on* (*plongeon* 'dive'), *-ure* (*brisure* 'splintering').

Among them, six exceptional and produce hapaxes. We have chosen to make a distinction between (low) productive affixation rules and affixes occurring in hapaxes. The latter are assigned a specific feature (see section 4.5).

> *-aire* (*commenter* 'comment'/*commentaire*),
> *-ande* (*offrir* 'offer'/*offrande*),

*-ange* (*louer* 'praise'/*louange*),
*-eur* (*errer* 'wander'/*erreur*),
*-ice* (*exercer* 'exercise'/*exercice*),
*-oire* (*interroger* 'question'/*interrogatoire*),

Some word-pairs in Verbaction and Lexeur are motivated by a clear (a strong) semantic relation, but their morphological relation is legitimate, because of it is imprecise or it involves a long-distance parenthood connection. For instance, Démonette has to record the *chromisation*/*chromer* 'chromization'$_{\text{N}}$/'chrome'$_{\text{V}}$ pair from Verbaction, where the *-ion* suffixed noun is derived from *chromiser* 'chromize' and the verb *chromer* is converted from the noun *chrome* 'chrome'$_{\text{N}}$. A further piece of information which has to be reported in Démonette is that *chromiser* and *chromer* are synonyms ( see Section 4.5).

## 4.2 Conversion

Conversion (or zero-derivation) names affixless derivation processes. When it involves a verb and a noun, the orientation is usually undecidable Tribout (2010), unless the stem of either the noun or the verb dictates otherwise. Three cases can be distinguished, when the relation between the verb $W_1$ and the noun $W_2$ is a conversion.

1. $W_1$ and $W_2$ have the same stem and the stem ends with a nominal suffix: e.g. *-ion*, in *addition*/*additionner* 'adding'$_{\text{N}}$/'add'$_{\text{V}}$ or *-ment*, with *réglement*/*réglementer* 'regulation'$_{\text{N}}$/'regulate'$_{\text{V}}$[3]. The noun is derived by suffixation, and therefore cannot be at the same time converted from the verb: in this case, $W_1$ is converted from $W_2$.

2. The noun stem corresponds either to a verbal past participle, e.g. *fait* 'fact'$_{\text{N}}$ related to *faire* 'do'$_{\text{V}}$, or to verbal Latin supine roots, e.g. *agrégat* 'aggregate'$_{\text{N}}$ related to *agréger* 'aggregate'$_{\text{V}}$ (about supine, cf. section 3.4): these stems can only originate from verbs. The conversion output is necessarily the noun, derived from a verb: here, $W_2$ is converted from $W_1$.

3. In all the other cases, no formal mean can help us decide whether the noun derives from the verb or vice-versa, e.g. with the relation between the verb *analyser* and the noun *analyse*.

[3]The *er* ending on the verb is the inflectional mark of infinitive. Therefore it is not a derivational suffix, nor does it belong to the verb stem

## 4.3 Cross-formation

Cross-formation defines regular morphological relations between two affixed words lacking a common ascendant. English word pairs like *pessimism*/*pessimist* belong to cross-formation paradigms. In our view, cross-derivation is a kind of predictable indirect relation between word pairs belonging to the same derivational family. In Verbaction and Lexeur, cross-formation concerns noun-noun pattern pairs, e.g. $X$`ion`/$X$`eur` (*imprécation*/*imprécateur* 'imprecation'/'imprecator'). The features Démonette assigns to the cross-formed word-pairs include regular definitions of each word with respect to the other, as shown in section 4.5

## 4.4 Back-formation

Back-formation (also called "substractive derivation") is traditionally defined as the process of creating a new word by removing an affix from its base (see Becker (1994) for a paradigmatic analysis of this phenomenon; see also Adams (2001); Nagano (2007); Shimamura (1983); Szymanek (2005) on back- and cross-formation). Back-formation is a diachronic process; for instance, English *orientate* is back-derived from *orientation* because the first appearance of the verb is more recent than that of the noun. Cases of back-formations in our corpus are observed with compound-like verbs (e.g. *hydroplaner* 'hydroplane', *radiodiffuser* 'broadcast') and their related suffixed action noun (resp. *hydroplanage* 'hydroplaning', *radiodiffusion* 'broadcasting'). Namer (2012) has provided evidence that these verbs can only be analysed as back-derived from the suffixed noun by analogy to the relation between the corresponding non-compound verb (e.g. *planer*, *diffuser*) and noun (resp. *planage*, *diffusion*). As will be shown below (section 4.5), the treatment of back-formation captures the fact that the relation is a classical but reversed suffixation process.

## 4.5 Strategies in Feature Assignation

With the newly integrated relations presented above contributed by Verbaction and Lexeur, the original organization of feature assignment in the first version of Démonette had to be improved, in order to capture the peculiarity of these phenomena (hapax formations, undecidable conversion orientation, long-distance relationships , cross- and back-formations), and to make them fit into existing paradigms. These relations are described by new combinations of features illustrated in Table 4, where the column heading **Cplx** stands for "complexity", **Orient**, for "orientation", $\mathbf{Exp_i}$, for "exponent" (of $W_i$) and **Def**, for "definition" of $W_1$ with respect to $W_2$. These features have all been presented in section 2

**Complexity:** The label lexical has been introduced to characterize the **complexity** feature (cf. Table 1) of hapax relations, which namely connect morphologically related words but not through a regular derivational relation. For instance, *mentir* 'lie'$_{\text{V}}$ and *mensonge* 'lie'$_{\text{N}}$ or *interrogatoire* 'questioning' and *interroger* 'question'$_{\text{V}}$ do not enter in any French derivational pattern, but they share enough meaning and formal properties to be considered as morphologically related (Table 4, rows 1 and 2).
The complex label is used to identify generation-skipping relationships, as for the *chromisation*/*chromer* pair in Table 4, row 3), where the noun, suffixed by *-ion* and the verb, derived by conversion (cf. Section 4.1) share the same nominal ancestor: *chrome*.

**Orientation:** This feature, used to indicates, when relevant, which of $W_1$ or $W_2$ descends from the other, is crucial for the distinction between conversion types (cf. Section 4.2), and for the identification and characterization of back-formation.
The differences between the three types of conversion are expressed in terms of **orientation** value: for a noun-verb pair (i.e. an entry $W_1 \leftarrow W_2$), the value of **orientation** is descending when the noun is derived from the verb (row 7 in Table 4), ascending when the noun is the base of the verb (row 8). The value is left blank when the conversion orientation is undecidable (row 9). In this case, both $W_1$ and $W_2$ can be analyzed as converted from each other (both $\mathbf{Type}_1$ and $\mathbf{Type}_2$ equal conv). Notice that when conversion is involved, $\mathbf{Exp}_i$ fields are not filled.
When it comes to back-formation (rows 10 and 11), the assigned features account for the fact that the noun (e.g. *radio-diffusion*) belongs to a derivational series (e.g. the class of *-ion* suffixed action nouns) and simultaneously serves as base for the verb (e.g. *radio-diffuser*). Compared to row 12, we can see that back-formation illustrated in row 10 differs only for its **orientation** value. The label indirect is used for word-pairs in a cross-formation relation, cf. rows 5 and 6: *imprécation* and *imprécateur* being equally complex and interpretable each with respect to the other, both $\mathbf{Type}_i/\mathbf{Exp}_i$ are filled, and the relation is symmetrical (compare rows 5 and 6). Finally, notice that when the relation is lexical, the value of **orientation** is left blank.

**Definition:** Irrespective to the orientation of the $W_1 \leftarrow W_2$ morphological relation, $W_1$ is always defined with respect to $W_2$, as shown in the last column of Table 4. There are two exceptions: words in a lexical relation (rows 1 and 2), and words in a formally complex relation but which lack mutual semantic motivation: as shown with the example of *syndicalism* 'syndicalism' / *syndiquer* 'syndicate' (row 13), the semantic distance between $W_1$ and $W_2$ is such that neither of them can be defined with respect to the other. On the other hand, other words connected by a complex

relation such as *chromisation* ($W_1$) and *chromer* ($W_2$), are assigned a definition with respect to each other, because of the synonymy between $W_2$ and the verb base of $W_1$ (here *chromiser*): here, since *chromer* has the same meaning as *chromiser* (i.e. "cover with chrome") , *chromisation*'s definition, with respect to *chromer* follows the same semantic pattern as with respect to *chromiser*, cf. row 4.

| | $W_1$ | $W_2$ | **Cplx** | **Orient** | $\mathbf{T_1/Exp_1}$ | $\mathbf{T_2/Exp_2}$ | **Def** |
|---|---|---|---|---|---|---|---|
| 1 | interrogatoire$_N$ 'questioning' | interroger$_V$ 'question' | lexical | – | suf/*-oire* | – | – |
| 2 | mensonge$_N$ 'lie' | mentir$_V$ 'lie' | lexical | – | suf/*-onge* | – | – |
| 3 | chromisation$_N$ | chromer$_V$ 'chrome' | complex | – | suf/*-ion* | conv | Action de chromer 'Action of chroming$_V$' |
| 4 | chromisation$_N$ | chromiser$_V$ 'chromize' | simple | – | suf/*-ion* | – | Action de chromiser 'Action of chromizing$_V$' |
| 5 | imprécation$_N$ 'imprecation' | imprécateur$_N$ 'imprecator' | simple | indirect | suf/*-ion* | suf/*-eur* | Action de l'imprécateur 'Action of the imprecator$_N$' |
| 6 | imprécateur$_N$ | imprécation$_N$ | simple | indirect | suf/*-eur* | suf/*-ion* | Agent de l'imprécation Agent of the imprecation$_N$' |
| 7 | agrégat$_N$ 'aggregate' | agréger$_V$ 'aggregate' | simple | descending | conv | – | Action de agréger 'Action of aggregating$_V$' |
| 8 | addition$_N$ 'adding' | additionner$_V$ 'add' | simple | ascending | – | conv | Action de additionner 'Action of adding$_V$' |
| 9 | analyse$_N$ 'analysis' | analyser$_V$ 'analyze' | simple | – | conv | conv | Action de analyser 'Action of analyzing$_V$' |
| 10 | radio-diffusion$_N$ 'broadcasting' | radio-diffuser$_V$ 'broadcast' | simple | ascending | suf/*-ion* | – | Action de radio-diffuser 'Action of broadcasting$_V$' |
| 11 | radio-diffuser$_V$ | radio-diffusion$_N$ | simple | descending | – | suf/*-ion* | Réaliser la radio-diffusion 'Perform the broadcast$_V$' |
| 12 | diffusion$_N$ 'spread' | diffuser$_V$ 'spread' | simple | descending | suf/*-ion* | – | Action de diffuser 'Action of spreading$_V$' |
| 13 | syndicalisme$_N$ 'syndicalism' | syndiquer$_V$ 'syndicate' | complex | indirect | suf/*-isme* | – | – |

Table 4: Identifying WF types by combining 'Complexity', 'Orientation', affixation 'Type' and 'Exponent', and 'Definition' values

# 5 Conclusion: Adaptability of Démonette to descriptive requirements

The examples of lexical data migration from Verbaction and Lexeur show that Démonette is able to accommodate new sort of information, of various nature and coming from sources whose primary purpose is not to describe the morphological structure of the entries that compose them. We have also seen that Démonette's architecture allows some of the fields to be left empty, depending on the nature of morphological relation connecting $W_1$ and $W_2$. Often values are provided —or computed during migration— to fill all the fields describing an entry $W_1 \leftarrow W_2$ in Démonette. But for some pairs, one or several fields are left blank: the orientation feature may be indeterminate; $W_1$'s definition is unfilled when $W_1$ cannot be spontaneously interpreted with regard to $W_2$.

Démonette is very flexible and it offers a unified representation for an —a priory— unlimited number of lexical resources having diverse content and purpose. The coverage of these resources is expanded by infering implicit morpho-semantic values from the input data. In this way, the resources find new uses and serve to weave an increasingly complex network within one same target structure, namely Démonette. In short, in order to ensure Démonette a long-lasting plasticity, the migration programs design has to make sure to favor the extension of its regular architecture without compromising its pre-existing structure, that is to allow for the coexistence of a set of core, fundamental features (connected words, parts-of-speech, morphological processes), with a set of significant —though not essential— ones (morpho-semantic, definitions), and a set of optional properties (stem graphical value, phonological representation).

The future developements of the Démonette database raise questions that further research will have to answer: How to extend the current set of morpho-semantic types, in order to accomodate new derivational relations, e.g. from new lexical sources? Which level of granularity level has to be chosen? Is it necessary to distinguish between property nouns, objective (*atomicité* 'atomicity') and subjective (*stupidité* 'stupidity') ones Koehl (2012), between properties referring to colours (*blondeur* 'blondness'), behaviours (*fourberie* 'deceit'), etc., or to separate true properties (*mortalité* 'mortality' meaning 'state of being mortal') from rates (*mortalité* meaning 'number of dead individuals')?

Other issues are related to the decision to connect or not a given ($W_1$, $W_2$), according to the semantic distance between $W_1$ and $W_2$: what formal criteria can be used to only include relevant indirect or complex relations, and exclude the more distant ones? The answer schould involve interpredictability, a notion formalized

for inflection by examining the statistical distribution of patterns of alternation and phonological shapes (see the seminal work of Ackermann et al. (2009), about the **Paradigm Cell Filling Problem**, as well as Bonami et al. (2011) and Ackerman & Malouf (2013)).

Finally, longer-term goals for further versions of Démonette will include its future capacity to combine information originating from different sources, especially extensive resources such as machine readable dictionaries such as GLAWI Sajous & Hathout (2015).

# References


Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89. 429–464.

Ackermann, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar*, 54–82. Oxford: Oxford University Press.

Adams, Valerie. 2001. *Complex words in english*. Harlow: Longman.

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

Becker, Thomas. 1994. Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of morphology 1993*, 1–25. Dordrecht: Kluwer Academic Publishers.

Bonami, Olivier, Gilles Boyé & Fabiola Henry. 2011. Measuring inflectional complexity: French and mauritian. In *Workshop on quantitative measures in morphology and morphological development*, .

Fabre, Cécile, Franck Floricic & Nabil Hathout. 2004. Collecte outillée pour l'analyse des emplois discordants des déverbaux en -eur.

Fellbaum, Christiane (ed.). 1999. *Wordnet: an electronic lexical database*. Cambridge, MA: MIT Press.

Hathout, Nabil. 2009. Acquisition of morphological families and derivational series from a machine readable dictionary. In Fabio Montermini, Gilles Boyé & Jesse Tseng (eds.), *Selected proceedings of the 6th décembrettes: Morphology in bordeaux*, Somerville, MA: Cascadilla Proceedings Project.

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

Hathout, Nabil. 2014. Phonotactics in morphological similarity metrics. *Language Sciences* 46. 71–83.

Hathout, Nabil & Cécile Fabre. 2002. Constitution et exploitation de lexiques de formes déverbales.

Hathout, Nabil & Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5). 125–168.

Hathout, Nabil & Ludovic Tanguy. 2002. Webaffix : Finding and validating morphological links on the WWW. In *Proceedings of the third international conference on language resources and evaluation*, 1799–1804. Las Palmas de Gran Canaria: ELRA.

Koehl, Aurore. 2012. *La construction morphologique des noms désadjectivaux suffixés en français*: Université de Lorraine dissertation.

Miller, Georges A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4). 335–391.

Nagano, Akiko. 2007. Marchand's analysis of back-formation revisited: back-formation as a type of conversion. *Acta Linguistica Hungarica* 54(1). 33–72.

Namer, Fiammetta. 2002. Acquisition automatique de sens à partir d'opérations morphologiques en farnçais : étude de cas. In *Actes de la 9e conférence annuelle sur le traitement automatique des langues naturelles (taln-2002)*, 235–244. Nancy: ATALA.

Namer, Fiammetta. 2009. *Morphologie, lexique et traitement automatique des langues : L'analyseur dérif*. Paris: Hermès Science-Lavoisier.

Namer, Fiammetta. 2012. Nominalisation et composition en français: d'où viennent les verbes composés ? *Lexique* 20. 173–205.

Namer, Fiammetta. 2013. A rule-based morphosemantic analyzer for French for a fine-grained semantic annotation of texts. In Cerstin Mahlow & Michael Piotrowski (eds.), *SFCM 2013* CCIS 380, 93–115. Heidelberg: Springer.

Namer, Fiammetta, Pierrette Bouillon, Evelyne Jacquey & Nilda Ruimy. 2009. Morphology-based enhancement of a French SIMPLE lexicon. In Nicoletta Calzolari, Anna Rumshisky, Pierrette Bouillon & Kyoko Kanzaki (eds.), *5th international conference on generative approaches to the lexicon*, 153–161. Pisa: ILC-CNR.

Rajman, Martin, Josette Lecomte & Patrick Paroubek. 1997. Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Tech. rep. EPFL & INaLF. GRACE GTR-3-2.1.

Sajous, Franck & Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, 405–426. Herstmonceux, UK.

Shimamura, Reiko. 1983. Backformation of english compound verbs. In John F. Richardson, Mitchell Marks & Amy Chukerman (eds.), *Papers from the parasession on the interplay of phonology, morphology and syntax*, 271–282. Chicago: Chicago Linguistic Society.

Szymanek, Bogdan. 2005. The latest trends in english word-formation. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 429–448. Dordrecht: Springer.

Tribout, Delphine. 2010. *Les conversions de nom à verbe et de verbe à nom en français*: Université Paris 7. Phd thesis.

Zeller, Britta D, Jan Snajder & Sebastian Padó. 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)*, 1201–1211. Sofia, Bulgaria.