

# 1

---

## Démonette, a French derivational morpho-semantic network

NABIL HATHOUT<sup>1</sup> & FIAMMETTA NAMER<sup>2</sup>

### Introduction

This paper addresses Démonette, a new lexical resource for French that offers derivational morpho-semantic information and whose content, design and development prospects make it a novel tool that fills a gap among lexical databases.

In its first version, Démonette contains 31,204 verbs, adjectives and nouns. It is a network, organized as a graph, in which vertices and edges carry morphological and semantic features; it is designed based on a cumulative conceptualization of the constructed meaning of derived lexemes and combines the semantic typing of the lexemes with the use of abstract definitions (i.e., generalized descriptions that allow us to group words with similar meanings). One principle underlying the construction of this resource is that the meaning of a derived word is a combination of partial semantic properties introduced by each of the derivational relations in which the word is involved. This principle has its foundation in the theoretical background we adopt, which considers that the application of Word Formation Rules (WFR) takes the form of a set of relations between words instead of a combination of morphemes.

Currently, Démonette's core structure and coverage arise from the contributions of two long-existing resources, following an assumption discussed in Hathout and Namer (2011): the DériF parsing system and

---

<sup>1</sup>CLLE/ERSS, CNRS & Université de Toulouse

<sup>2</sup>Université de Lorraine & ATILF – UMR 7118

the Morphonette network. On the one hand, these resources are built on conceptually opposite frameworks, but on the other hand, they are complementary in terms of coverage and share theoretically consistent linguistic information. Nevertheless, as we see in this paper, Démonette has a distinctive formal architecture that is enriched by a set of original semantic features, and therefore, it is far from being simply a merger of DériF’s derivational analysis and Morphonette’s paradigms.

Indeed, Démonette has two major objectives. The first addresses theoretical issues, with the intent of constructing a large-scale, reliable and linguistically sound derivational network for French that is in accordance with the principles of a modern vision of Word Formation, i.e., word-based morphology. The second concerns NLP and linguistic engineering, which will benefit from the various morphological and morpho-semantic information Démonette provides.

The remainder of the paper is organized as follows. After a review of comparable resources in Section 1.1 and a brief presentation of the major features of Démonette, Section 1.2 describes DériF and Morphonette, the resources from which Démonette originates. Then, Section 1.3 outlines Démonette’s underlying model and explains how information is extracted from these resources to fuel this new lexical resource. Next, Section 1.4 details the cumulative conceptualization of constructed meaning, the typing of derived lexemes and the use of abstract and concrete definitions. A qualitative and quantitative characterization of the current version of Démonette is then provided in Section 1.5. Finally, in Section 1.6, we discuss various issues related to the extension of Démonette and the possibility of discovering and characterizing new derivational relations that could be incorporated into this new resource without relying on DériF or Morphonette.

## 1.1 State of the Art

Démonette features a unique combination of traits: it is a computational morphology resource, it is dedicated to derivational morphology, it is relation-oriented, it includes lexical semantic information, and its semantic descriptions are cumulative and involve semantic types and both concrete and abstract definitions. Other resources and tools share some of these characteristics.

The most common morphological resources are inflectional lexicons. Several large inflectional lexicons exist for French, such as *Lefff* (Sagot et al., 2006), which also includes syntactic subcategorization frames; *Morphalou* (Romary et al., 2004) or *TLFnome*, which were created from the *Trésor de la Langue Française* word list, and *GLAFF*, which was

extracted from the French Wiktionary and is notable for its significantly larger size and phonemic transcriptions (Sajous et al., 2013, Hathout et al., 2014).

In addition to these inflectional resources, some derivational morphological resources are available, though much fewer in number. The best known is certainly CELEX, which addresses English, German and Dutch (Baayen et al., 1995). It provides a comprehensive description of these three lexicons that includes phonology, inflectional and derivational morphology, syntactic categories and corpus frequency. CELEX adopts a conventional model of derivational morphology: lexemes have morphological structures that are represented as morpheme decompositions, as in (1) for the noun *occupation*. CELEX contains no lexical semantic relations and provides no characterization of word meaning.

(1) ((occupy) [V], (ation) [N|V.]) [N]

There are also more modest derivational resources that essentially provide morphologically related word pairs, such as CatVar (Habash and Dorr, 2003) for English or DerivBase for German (Zeller et al., 2013). A more sophisticated resource, Morphonette (Hathout, 2011), is described in detail in Section 1.2.3. Note that more specialized resources have also been created for French, such as Verbaction (Tanguy and Hathout, 2002; see Section 1.5).

Work related to the construction of derivational morphological resources is relatively scarce. Most efforts in computer morphology have focused on the development of morphological analyzers (see (Hammarström and Borin, 2011) for a summary). Many have been developed over the past two decades to meet the needs of morphological analysis in NLP and to overcome the limitations of the lexicons. Such parsers split words into morphemes. Some of them add labels to identify morpheme variants. They are usually based on the Harrisian model (Harris, 1955) and on a variety of machine learning methods. Therefore, they are relatively language independent, although many are designed to yield better results for Western European concatenative languages such as English, German and French. These systems include Linguistica (Goldsmith, 2001), Morfessor (Creutz, 2003, Creutz and Lagus, 2005) and Bernhard's (2006) parser. In this respect, this trend is similar to that observed in POS-tagging and syntactic parsing. Other systems, grounded in more recent morphological theories have also been developed, such as DériF (Namer, 2009), which is presented in Section 1.2.2 (for a general overview, see (Bernhard et al., 2011, Namer, 2013b)).

Morphological descriptions may also be found in certain general lexical semantic resources, such as wordnets (Fellbaum, 1999) or Jeux de

Mots (Lafourcade and Joubert, 2013). Princeton WordNet has a morphosemantic supplement that exhibits several similarities with Démonette (Fellbaum and Miller, 2003). This “over-layer” is a lexicon of 17,740 word pairs. The first word in the relation is characterized by a semantic type (event, condition, income, property, location, etc.). In addition, the meaning of each word in each pair is described by a definition (i.e., the gloss of the synset it belongs to). Therefore, these definitions do not always reference the meaning of the other word of the pair. Similar morphosemantic descriptions have also been added to the Czech and Turkish wordnets (Pala and Hlaváčková, 2007, Bilgin et al., 2004). Note also that the structure defined by such a supplement does not exactly coincide with the organization of WordNet into synsets because the morphosemantic relations hold only between the words that make up the synsets and not between synsets, as for the other lexical semantic relations.

Jeux de Mots is large lexical network for French that was developed using an online game with a purpose. It describes a wide range of classic lexical relations (synonymy, hypernymy, meronymy, etc.). This resource also contains derivational morphological relations in the following forms: verb-process (*construire* ‘build’ → *construction* ‘construction’; 6453 relations r\_verbe-action) or process-verb (*jardinage* ‘gardening’ -> *jardiner* ‘gardenv;’ 6388 relations r\_action-verbe). Because it was created by online game players, Jeux de Mots has very broad lexical coverage but lacks a systematic treatment of the encoded lexical and morphological relations, especially the latter.

The definitions found in WordNet are concrete (i.e., fully instantiated) definitions (see Section 1.3.4). Other resources, such as FrameNet, feature abstract definitions in which some of the defining elements are variables. FrameNet consist of lexical units (words) associated with semantic frames and exemplified by annotated examples. A “semantic frame [is] a script-like conceptual structure that [jointly] describes a particular type of situation, object, or event along with its participants and properties” (Ruppenhofer et al., 2006). A frame includes both a definition of the concept and its potential participants. These definitions are abstract to the extent that participants are represented by variables that indicate their semantic roles. Furthermore, the variables are typed, as in Démonette (see Figure 1). The definition of a participant is relative to both the situation and the other participants. In Démonette, every elementary relation is described separately. Moreover, FrameNet’s definitions are less formal than those of Démonette. In particular, we observe some variation in the expression of the predicate in Figure 1: it is represented by “cut” in the Agent definition and

“slicing” in the **Pieces** definition.

### Cutting

*Definition:* An **Agent** cuts an **Item** into **Pieces** using an **Instrument** (which may or may not be expressed).

*Frame elements:*

*Core:*

Agent      The **Agent** is the person cutting the **Item** into **Pieces**.

*Semantic Type:* Sentient

---

Item	The item that is being cut into <b>Pieces</b> .
------	---

---

Pieces	The <b>Pieces</b> are the parts of the original <b>Item</b> that are the result of the slicing.
--------	---

*Non-core:*

Instrument      The **Instrument** with which the **Item** is being cut into **Pieces**.

*Semantic Type:* Physical\_entity

---

Manner	<b>Manner</b> in which the <b>Item</b> is being cut into <b>Pieces</b> .
--------	--

---

*Semantic Type:* Manner

---

...

FIGURE 1 Excerpt of the *cutting* semantic frame

## 1.2 Data

Unlike the morpho-semantically annotated lexical resources presented above, Démonette is a lexical network that is designed to achieve a quadruple objective:

1. Connect the members of a derivational family (i.e., words that share the same stem), be they in a direct relationship (e.g., between the noun *essorage* ‘spinning’ and its verbal base *essorer* ‘(to) spin’) or an indirect relationship (e.g., between the nouns *essorage* and *essoreuse* ‘spin-dryer,’ which share the same base). These connections are all bi-oriented.
2. Label each relation with a definition that describes the meaning of one of the connected words with respect to that of the other. For instance, the meaning of *essorage* is described as the ‘action of *essorer*’ or as the ‘action performed by an *essoreuse*;’ likewise, *essorer* is defined as ‘(to) perform *essorage*.’
3. Provide the words in the network with morphological and semantic tags. For instance, Démonette indicates that *essorage* is an

action noun (semantics) derived from a verb (morphology) and formed with the *-age* suffix (morphology).

4. Bring together, in the form of abstract notations, the definitions of words that share similar semantic features. For instance, the definition of *essorage* with respect to *essorer* can be regarded as an instance of the abstract expression ‘action of PREDICATE,’ which subsumes all verb-based action nouns in the network, regardless of their morphological properties.

To achieve these multiple goals, three sources of information have been used, each of which contributes in its own way to the content of Démonette. These sources are as follows: a list of words; the morphological analyzer DériF, as the primary source of morpho-semantic information; and the Morphonette lexicon, which contains a number of indirect relations.

### **1.2.1 Démonette’s list of words**

The selection of the words included in the current version of Démonette was guided by two criteria. The first was the alignment of the initial resources to allow the morphological graph to be created using a single set of predefined vertices. This set of words was extracted from the TLFnome lexicon<sup>3</sup>. The reason for this choice was twofold: DériF has primarily been developed based on the TLFnome lexicon (cf. Section 1.2.2), and Morphonette has been built from this lexicon (cf. Section 1.2.3).

Achieving good convergence between resources as different as DériF and Morphonette requires the resolution of certain formal and practical implementation issues. To this end, the initial set of words was reduced to lexemes constructed using seven suffixes and their possible verb bases (see Section 1.3.1).

### **1.2.2 DériF**

DériF (Namer, 2009, 2013b) is a morphological analyzer that implements WFRs developed by linguists. It has two major features:

1. The analyses are controlled by a set of exceptions that account for some of the irregularities that have accumulated during the evolution of the lexicon.
2. It provides each derived word with a (concrete) definition, that is, a phrase that expresses its morphologically constructed mean-

---

<sup>3</sup>TLFnome is a lexicon created from the TLF word list. It contains 97,000 lemmas and is 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/](http://www.cnrtl.fr/lexiques/morphalou/).

ing with respect to its base when the word is formed through derivation, as in (2a), or with respect to its bases in the case of compounding, as in (3a). These definitions are reminiscent of those of WordNet (Miller et al., 1990, Fellbaum, 1999).

DériF analyzes POS-tagged lemmas. It recursively applies the WFRs 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 (2b) and (3b). The first element in the list is the analyzed word.

- (2) 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) a. *fluidiforme/ADJ* ← *fluide/ADJ, forme/NOM*  
*dont la forme est fluide*  
‘whose shape is fluid’
- b. *(fluidiforme/ADJ, [fluide/ADJ], forme/NOM)*  
‘(fluidiform, [fluid], form)’

When several competing derivations are possible, for example, in the case of a bracketing paradox (Spencer, 1988, Becker, 1994), DériF chooses the solution that is best motivated linguistically. For example, *désenneigement* ‘snow clearance,’ could potentially be derived from *enneigement* via prefixation with *dé-* or from *désenneiger* ‘clear the snow from’ via suffixation with *-ment*. The base *désenneiger* is preferred to *enneigement* because *dé-* prefixation constructs verbs rather than nouns (Amiot, 2008), so DériF analyzes *désenneigement* as in (4) and not (5). In the remainder of this section, DériF’s analyses are presented as lists of antecedents.

- (4) *désenneigement/NOM* : *(désenneigement/NOM, désenneiger/VER, enneiger/VER, neige/NOM)*
- (5) *désenneigement/NOM* : *(désenneigement/NOM, enneigement/NOM, enneiger/VER, neige/NOM)*

However, when multiple analyses are possible, all solutions are provided. This is the case in (6), where the order of rule application (6a vs 6b) does not matter, and in (7), where *importable* is polysemous, as it can mean either ‘unwearable’ or ‘importable.’

- (6) a. *contre-manifestation/NOM* : (*contre-manifestation/NOM, manifestation/NOM, manifester/VER*)  
‘(counter-demonstration, demonstration, demonstrate)’
  - b. *contre-manifestation/NOM* : (*contre-manifestation/NOM, contre-manifester/VER, manifester/VER*)  
‘(counter-demonstration, counter-demonstrate, demonstrate)’
- (7) a. *importable/ADJ* : (*importable/ADJ, portable/ADJ, porter/VER*) ‘(unwearable, wearable, wear)’
  - b. *importable/ADJ* : (*importable/ADJ, importer/VER, import/NOM*) ‘(importable, import<sub>V</sub>, import<sub>N</sub>)’

By design, DériF is also capable of analyzing neologisms. The rule system predicts their bases and calculates their morphologically constructed meanings based on their morphological structures and their POSs (in (8), the noun *schtroumpfement* ‘smurfing’ is formed via suffixation with *-ment*).

- (8) *schtroumpfement/NOM* : (*schtroumpfement/NOM, schtroumpfer/VER*) ‘(smurfing<sub>N</sub>, smurf<sub>V</sub>)’

DériF accounts for lexicalization, particularly for words that have morphologically complex forms but a meaning that cannot be predicted (anymore) from these forms. This is managed through exception lists that block rule application. For example, the inclusion of *pension* ‘pension’ among the exceptions prevents the system from analyzing it as the deverbal action noun of *penser* ‘think’ (9a), following the model of *pression* ‘pressure’ derived from *presser* ‘press’ (9b).

- (9) a. *pension/NOM* : (*pension/NOM*)
- b. *pression/NOM* : (*pression/NOM, presser/VER*)

DériF assigns to each analyzed word a concrete definition (see, e.g., (10), where the logical “OR” indicates an ambiguous meaning). It also provides a set of features that reflect the constraints imposed by the morphological rules (Namer, 2002, Namer et al., 2009). For instance, the verb-to-noun *-eur* suffixation rule predicts the assignment of lexical knowledge: derivatives are concrete and countable nouns; bases are dynamic, agentive verbs (11).

- (10) *chanteur/NOM* : (*agent masculin habituel OR auteur masculin exceptionnel OR instrument*) de chanter  
‘singer<sub>(M)</sub>: (usual masculine agent OR exceptional masculine agent OR instrument) of singing’<sup>4</sup>

---

<sup>4</sup>Subscripts (M) (resp., (F)) indicate that the noun is masculine (resp., feminine).

- (11) a. *chanter/VER*: [ aspect = *dynamique*, sous\_cat = <NPagent, ...> ]  
           ‘[ aspect = dynamic, sub-cat = <NPagent, ...> ]’  
 b. *chanteur/NOM*: [ concret = *oui*, comptable = *oui* ]  
       ‘[ concrete = yes, countable = yes ]’

### 1.2.3 Morphonette

Morphonette is a French lexical network based on a relational and paradigmatic conceptualization of morphology. The morphological properties of lexemes are described by the paradigms to which they belong. For example, the properties of a derivative such as *modifiable* ‘modifiable’ can be minimally described in terms of its derivational family, which encompasses the lexemes in (12a), and its derivational series, which contains all *-able* derivatives (12b).

- (12) a. *modifier/VER*, *modification/NOM*, *modificateur/NOM*,  
*modificatif/ADJ*, *modifiant/ADJ*, *modifieur/NOM*,  
*immodifiable/ADJ*, etc.  
       ‘modify, modification, modifier, amending, modifying,  
       modifier, unmodifiable, etc.’
- b. *agaçable/ADJ*, *agitable/ADJ*, *chevauchable/ADJ*,  
*définissable/ADJ*, *différenciable/ADJ*, *rechargeable/ADJ*,  
*réconciliable/ADJ*, *soutenable/ADJ*, etc.  
       ‘irritable, moveable, rideable, definable, differentiable,  
       refillable, reconcilable, sustainable, etc.’
- (13) (*modifiable/ADJ*, *modificateur/NOM*,  
       {*amplifiable/ADJ*, *glorifiable/ADJ*, *identifiable/ADJ*,  
       *justifiable/ADJ*, *clarifiable/ADJ*, *mystifiable/ADJ*,  
       *rectifiable/ADJ*, *sanctifiable/ADJ*, *simplifiable/ADJ*,  
       *spécifiable/ADJ*, *unifiable/ADJ*, *vérifiable/ADJ*})  
       ‘(modifiable, modifier,  
       {amplifiable, glorifiable, identifiable, justifiable, clarifiable,  
       mystifiable, rectifiable, sanctifiable, simplifiable, specifiable,  
       unifiable, verifiable})’

As stated above, Morphonette was constructed from TLFnome. It is composed of **filaments**, i.e., triplets of the form  $(w, r, s_r(w))$ , 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  $r$ ; in other words,  $s_r(w)$  is the set of words that participate in relations similar to the relation between  $w$  and  $r$ . Thus, a word  $u$  belongs to  $s_r(w)$  if there exists a word  $v$  such that  $w : r = u : v$  (i.e., such that  $w$ ,  $r$ ,  $u$ , and  $v$  form an analogy).

Example (13) shows the filament of the adjective  $w = \text{modifiable}$  ‘modifiable’ for  $r = \text{modificateur}$  ‘modifier’; *amplifiable* belongs to the series of this filament because of the analogy *modifiable:modificateur=amplifiable:amplificateur*.

In Morphonette, an entry has as many filaments as there are members in its derivational family. Some filaments overlap considerably. Others describe different properties of the same word. For example, *travailleur* ‘worker<sub>(M)</sub>’ has in its family the noun *travailleuse* ‘worker<sub>(F)</sub>’ and the verb *travailler* ‘work.’ Its series with respect to the verb contains derivatives such as *ravageur* ‘ravager’ and *cisailleur* ‘shearer,’ which can be associated with the verbs *ravager* ‘ravage’ and *cisailler* ‘shear,’ respectively. Moreover, the series of *travailleur* relative to *travailleuse* contains masculine nouns such as *deuilleur* ‘mourner’ or *volailleur* ‘poulterer,’ which have feminine counterparts constructed with /ɔz/ (*deuilleuse* and *volailleuse*, respectively) but no base verb. In fact, *deuilleur* derives from the noun *deuil* ‘mourning,’ and *volailleur* derives from the noun *volaille* ‘poultry.’ Thus, *travailleur* is a member of two distinct series, one corresponding to the property of being a deverbal noun and the other to that of being associated with a feminine noun ending in /ɔz/.

The objective of Morphonette is to discover and represent all derivational relations between TLFnome entries (Hathout, 2011). Their identification is conducted in three steps.

- (a) Select, for each entry, a neighborhood consisting of the 100 most morphologically similar words. This neighborhood is defined by the Proxinette measure (Hathout, 2009, In press). The similarity of two words is estimated with respect to the number and specificity of the formal features (*n*-grams of characters) they share. Proxinette is comparable to the distance of De Pauw and Wagacha (2007).
- (b) Collect all formal analogies that exist between an entry and its neighbors (Lepage, 1998, Stroppa and Yvon, 2005). Analogies are identified on the basis of editing signatures calculated for each pair  $(w, u)$ , where  $w$  is an entry and  $u$  is one of its neighbors (Lepage, 2004, Gosme and Lepage, 2009). Pairs that share the same signature form analogies.
- (c) Apply three criteria to separate families from series and to select the quadruplets that are more likely to be morphologically valid.

- The first is a categorical criterion that separates relations between family members from relations between series members in analogies in which the four words do not belong to the same category.
- The second criterion is related to the size of the series: two words that belong to the same derivational family typically participate in a large number of analogies. Filaments with series of fewer than 10 words are discarded.
- The third criterion is based on the clustering coefficient of the series (Watts and Strogatz, 1998): elements of the series of a given filament tend to appear together in the series of other filaments. The clustering coefficient for the series members is calculated (i.e., the ratio between the number of triangles and the number of triples that contain these members). Series members with a clustering coefficient of less than 0.66 are removed.

These three criteria, in combination with the restrictions imposed during the first two stages, allow us to obtain a more reliable network, although a relatively small one. It includes 29,310 words, 96,107 filaments, 96,107 relations between members of the same derivational family and 1,160,098 relations between members of the same derivational series. Filaments are interesting because they describe the relations between derivatives and their bases (direct relations), the members of their derivational families (indirect relations) and the members of their derivational series.

#### 1.2.4 Conclusion

As will be seen in the next section, the general structure of Démonette that we outlined at the beginning of Section 1.2 enables us to organize the morphological information produced by DériF’s analyses or contained in Morphonette in a novel fashion. Two facts must be borne in mind: First, only a part of the available information has been integrated into this current version of Démonette because our priority has been to create the network and resolve the difficulties involved in making the two input resources converge. In other words, the current version of Démonette serves as a proof of concept. Second, we will demonstrate that a significant proportion of the features provided by Démonette yield new pieces of information.

### 1.3 Principles and Methods

This section is devoted to the in-depth presentation of the principles governing the internal organization of the Démonette lexical

network, the required morpho-lexical information, and the conditions under which that information is produced. Afterward, we will see the role that the resources described in the previous section play in the achievement of these objectives and present the knowledge directly calculated for Démonette.

### 1.3.1 Démonette's general principles

Démonette features a novel architecture among morphological resources. It is a lexical network that is implemented as a directed graph, which incorporates the outcomes of existing resources. Its goal is ultimately to provide bi-oriented definitions for each pair of related words such that, in the resulting lexicon, words are defined with respect to each of the words with which they have a morphological relation: derived lexemes are connected to their bases by directed edges that we call **direct relations** and to the other members of their derivational families by directed edges that we call **indirect relations**. For ex-

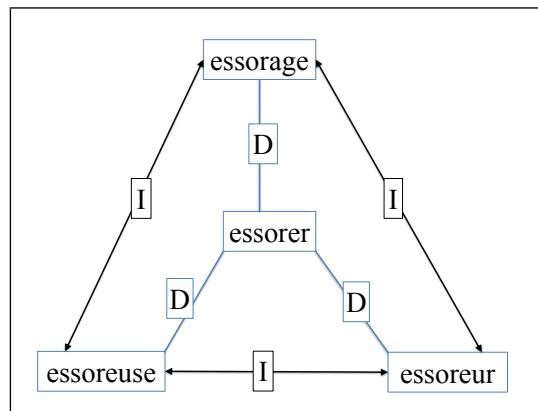


FIGURE 2 Direct and indirect relations in the *essorer* derivational family

ample, the edges listed in (14) and illustrated in Figure 2 connect the members of the family of the verb *essorer* ‘spiny’: *essorer:essorage* ‘spiny:spinning<sub>N</sub>’ is a (D)irect relation (typeset in bold in (14)), whereas *essorage:essoreur* ‘spinning<sub>N</sub>:wringer’ and *essorage:essoreuse* ‘spinning<sub>N</sub>:spin-dryer’ are (I)ndirect relations.

- (14) ***essorage:essorer***; *essorage:essoreur*; *essorage:essoreuse*;  
***essoreur:essorer***; *essoreur:essorage*; *essoreur:essoreuse*;  
***essoreuse:essorer***; *essoreuse:essorage*; *essoreuse:essoreur*.

In its current version, Démonette’s coverage includes suffixed nouns and adjectives from TLFnome and their bases. The derivatives are action nouns ending in *-age*, *-ment* and *-ion*; agent nouns

ending in *-eur*, *-euse* and *-rice*; and property adjectives ending in *-if*. We call these sets of derivatives **derivational series** (see Table 1). In Section 1.4.2, we explain why we focus on these suffixes. Derivational series are potentially infinite sets of words that have the semantic and formal (i.e., phonological) features that characterize words derived through a WFR. The members of a derivational series are typically all derived through the same WFR and are defined in the same manner with respect to their respective bases. Consider, for instance, the noun *conductrice* ‘driver<sub>(F)</sub>,’ from the derivational series that includes derivatives constructed via the verb-to-noun *-rice* WFR. Its definition with respect to the base verb is *Agent féminin habituel OR auteur féminin exceptionnel OR instrument*) de conduire ‘Usual feminine agent OR exceptional feminine agent OR instrument of the action of driving.’ The definition of the other members of its derivational series follow exactly the same pattern, e.g., *contestatrice* ‘protester<sub>(F)</sub>’ is defined as *Agent féminin habituel OR auteur féminin exceptionnel OR instrument*) de contester ‘Usual feminine agent OR exceptional feminine agent OR instrument of the action of contesting.’ Démonette does not yet contain edges between derivatives constructed via the same affixation. For example, there is no edge in the graph to connect *essorage* with *séchage* ‘drying<sub>N</sub>.’ Démonette connects only words that belong to the same **morphological family**. A derivational family is a set that contains all words that share (any formal variant of) the same root. In its current version, the Démonette network is a non-connected graph that contains a single isolated subgraph for each derivational family.

formal features	semantic features	categorical relation
<i>-age, -ment, -ion</i>	action	V → N
<i>-eur, -euse, -rice</i>	agent	V → N
<i>-if</i>	property	N, V → A

TABLE 1 Selected suffixes

Each vertex that represents a lexeme carries a certain set of features. These are the lexeme’s POS and morphosyntactic features and, possibly, the type of affixation used for its construction. For example, the features of the vertex representing *essorage* indicate that it is a masculine noun constructed via suffixation with *-age*.

Thus, Démonette offers a detailed morphological representation of the set of lexemes it describes.

As stated above, Démonette also provides lexemes with descriptions of their “morphological meanings.” Such a description builds upon the graph edges. For a given vertex, it comprises (*i*) its semantic type and (*ii*) a set of **definitions** that describe the semantic relations supported by the morphological relations of the lexeme (see Section 1.4.3). For example, the derivational relation that connects *essorer* to *essorage* carries two concrete definitions, which describe the contribution of the morphological relation to the meanings of the two words in semi-natural language, e.g., (15)). These definitions are called **concrete** because they are fully instantiated. The definition (15a) is called **oriented** (base → derivative) and describes the meaning of *essorage* relative to that of *essorer*. The second one (15b) is said to be **opposite** (derivative → base) and describes the meaning of *essorer* relative to *essorage*.

- (15) a. (*action OR résultat de l'action*) *de essorer*  
‘(action OR result of the action) of spinning’
- b. *réaliser l'essorage* ‘perform the spinning’

These definitions follow from the morphological relations. One consequence of this conceptualization of meaning is that a lexeme that participates in several (direct or indirect) morphological relations may potentially have several definitions (see Section 1.4.1). In addition to concrete definitions (15), edges also carry both oriented and opposite **abstract** definitions, in which the meanings of the related words are represented by the semantic types of the corresponding lexemes. For example, *essorage* is typed as an “action noun,” with the label @ACT. The abstract definitions presented in (16) correspond to the concrete definitions in (15).

- (16) a. (*action OR résultat de l'action*) *de @*  
‘(action OR result of the action) of @’
- b. *réaliser le / la @ACT* ‘perform the @ACT’

These abstract definitions are obtained by replacing *essorer* in (15a) and *essorage* in (15b) with their **semantic types**, that is, with @ (for “predicate”) and @ACT (for “action noun”), respectively (see below). These definitions place *essorage* in the class of action nouns, or nouns that denote process results. These semantic types are therefore induced by the relations in which the lexemes are involved.

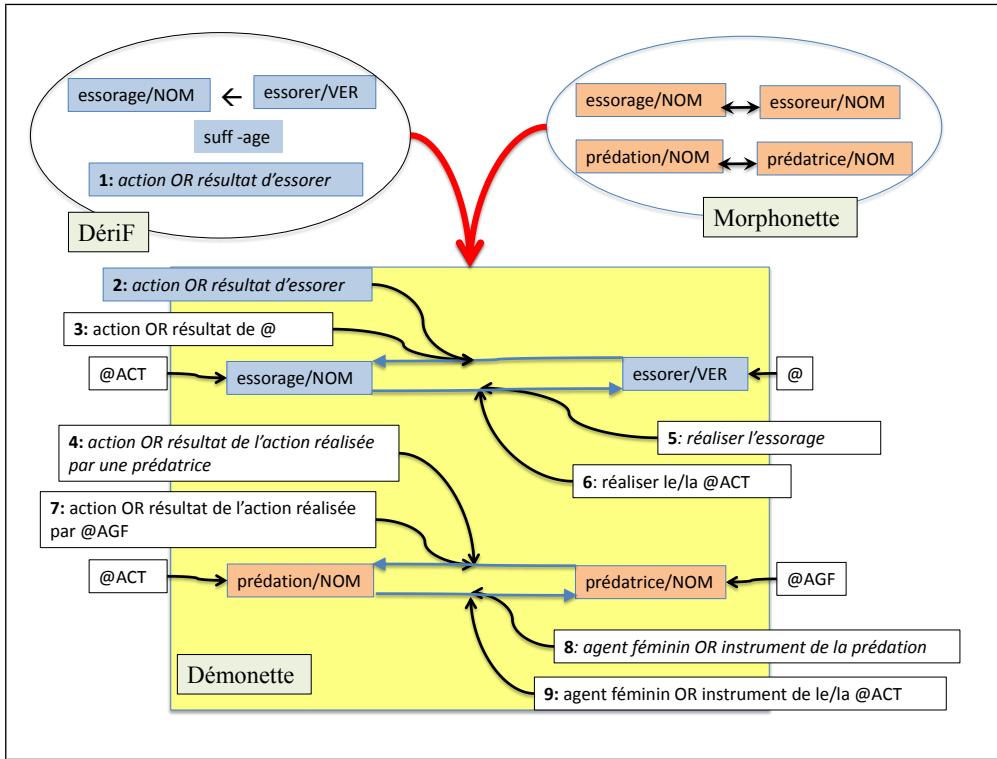


FIGURE 3 Schematic illustration of Démonette content

Figure 3 gives an overview of Démonette’s content and summarizes the manner in which the contributions from DériF and Morphonette are used (represented by red arrows). The nouns *prédatation* and *prédatrice* mean ‘predation’ and ‘female predator,’ respectively. With respect to the latter, the former can be defined as an “action OR result of the action realized by a predator(F).”

Correspondingly, a *prédatrice* is a “feminine agent OR instrument of predation.” As will be detailed below, DériF provides base-derivative relations and concrete oriented definitions (e.g., 1 and 2, in blue rectangles), whereas the input from Morphonette introduces indirect relations (orange rectangles). Finally, new features are calculated by Démonette (white rectangles): concrete opposite definitions (e.g., 5) and concrete cross-definitions (e.g., 4 and 8); semantic types, abstract oriented definitions (e.g., 3), abstract opposite definitions (e.g., 6) and abstract cross-definitions (e.g., 7 and 9).

The motivations behind the choice of semantic types on the one hand and opposite and cross-definitions on the other hand will be explained in Sections 1.4.2 and 1.4.3. First, we will explain how relevant features are imported from DériF and Morphonette (Sections 1.3.2 and 1.3.3). Then, in Section 1.3.4, we will provide an overview of the scripts that

have been implemented to generate Démonette's original information, i.e., the contents of the white boxes in Figure 3.

It is worth noting that morpho-semantic definitions in Démonette describe relations between the parts of word meanings that are used or constructed via morphology. Each of these parts corresponds to a base meaning (i.e., the part of the meaning of a base that is used to construct the meaning of a derivative) or a constructed meaning (i.e., the meaning attributed to a derivative by a morphological construction). Therefore, morpho-semantic definitions are not intended to describe all the semantic nuances of base and derived lexemes. This restriction can significantly reduce the inventory of morpho-semantic relations. It also limits the number of semantic types, as they are directly determined by these relations. For example, Fellbaum et al. (2009) list 9 classes of morpho-semantic relations for the *-er* suffixation in English, which all are also associated with other morphological constructions. For instance, instrument nouns are formed through *-er* suffixation (e.g., *slipper*) but also through compounding with *-graph* (*cardiograph*) or *-scope* (*fiberscope*). Likewise, event and result nouns may be suffixed via *-ion* (*detection*), *-ment* (*enrichment*) or *-ery* (*recovery*), among others. In a similar study of Czech, Pala and Hlaváčková (2007) use 16 classes of morpho-semantic relations such as property nouns, possession relation and gender change. We estimate that eventually, fewer than thirty relation classes will enable us to describe 95% of the direct morpho-semantic relations in the lexicon.

### 1.3.2 Feeding Démonette with information extracted from DériF

As illustrated in Figure 3, we used DériF's analyses for two purposes: (i) to create part of Démonette's graph structure — the set of vertices and edges — and (ii) to calculate the attributes of these vertices and edges.

(i) We first ran DériF on TLFnome and created an initial derivational graph whose vertices were TLFnome entries and whose edges were the direct derivational relations between the derived words and their direct antecedents (bases or compounding elements). We then identified the vertices that belonged to the Démonette list of words (as defined in Section 1.2.1). This identification was performed in 3 stages. We first selected the nouns formed via one of the six verb-to-noun suffixes of Table 1. In the second step, we added the base verbs of the derived nouns into the Démonette word list. We then added any adjective ending in *-if* whose base was either a noun or a verb already included in the list. Finally, we reduced the initial graph to this list of vertices,

identified the derivational families in the reduced graph and added indirect relations between their members. The resulting graph structure was directly integrated into Démonette.

(ii) In addition, oriented concrete definitions computed by DériF, such as the one connecting the meaning of the noun *essorage* to that of its base verb *essorer* (cf. (15a)), were associated with the corresponding derivative-base edges in Démonette.

### 1.3.3 Feeding Démonette with information extracted from Morphonette

Morphonette's data could not be directly incorporated into Démonette despite the fact that Morphonette itself is already a morphological network. We were first obliged to reconstruct pairs of lexemes to identify their morphological properties. For this purpose, 80,115 pairs of lexemes sharing the same stem were reconstructed from derivational families, as in (17). Pairs of lexemes participating in the same morphological relation were grouped together, such as those formed as a noun via *-euse* and an adjective via *-if* (18).

- (17) (*brouilleur/NOM, brouillage/NOM*)  
‘(scrambler, scambling)’
- (18) *collectif/ADJ:collectionneuse/NOM =*  
*dissertatif/ADJ:disserteuse/NOM =*  
*perfectif/ADJ:perfectionneuse/NOM =*  
*portatif/ADJ:porteuse/NOM =*  
*possessif/ADJ:possesseuse/NOM =*  
*sélectif/ADJ:sélectionneuse/NOM*  
‘*collective:collector<sub>(F)</sub> = discursive:speaker<sub>(F)</sub>* =  
*perfective:improver<sub>(F)</sub> = portable:carrier<sub>(F)</sub>* =  
*possessive:possessor<sub>(F)</sub> = selective:selector<sub>(F)</sub>*’

Thus, 14 sets of pairs were obtained, as in (18), to represent all possible combinations  $X_1 \text{suff}_1 : X_2 \text{suff}_2$ , where  $\text{suff}_1$  and  $\text{suff}_2$  are two of the seven suffixes from Table 1 (see Section 1.3.1) and stems  $X_1$  and  $X_2$  are sufficiently similar to be regarded as regular variants. These sets of pairs were added to Démonette as indirect relations, and the suffixes were used to assign semantic types to the vertices.

### 1.3.4 Démonette's original knowledge

As demonstrated above, DériF and Morphonette provided various types of information for the construction of Démonette:

- Démonette vertices;
- Démonette edges;

- word formation rules;
- direct relations between derived words and their bases;
- indirect relations between derived words sharing the same stem; and
- oriented concrete definitions carried by the direct relations.

Démonette entries are pairs of lexemes in the form  $(L_1, L_2)$  that are related via a direct or indirect relation. To assign these entries their specific Démonette features, we first match them to annotation templates that were manually designed in accordance with linguistic criteria. We then apply three scripts:

**Semantic type assignation.** The semantic types of  $L_1$  and  $L_2$  are assigned based on their morphological structures and parts of speech. For instance, the conditional statement (19) indicates that nouns ending in *-eur* that are related to a verb are masculine agents:<sup>5</sup>

```
(19) if (L1.cat = "NOM" and L1.structure = "Xeur"
      and L2.cat = "VER")
      then L1.type = "@AGM"
```

**Concrete opposite and abstract definitions.** The templates for concrete and abstract opposite definitions depend on the corresponding oriented definition. For instance, if  $L_2$  is a noun ending in *-age* and  $L_1$  is a verb, then the  $L_2$  definition corresponds to the schema “Action OR result of the action of  $L_1$ ” and the  $L_1$  definition corresponds to the schema “Perform the  $L_2$ .<sup>6</sup>” The assignment is implemented via statement (20a) for a concrete opposite definition and via (20b) for an abstract opposite definition.

```
(20) a. if (L1.cat='NOM' and L1.structure = Xage
           and L2.cat = 'VER'
           and L1.def = "Action OR result of the action of L1.value")
           then L2.def = "Perform the L2.value"
       b. if (L1.cat='NOM' and L1.structure = Xage
           and L2.cat = 'VER'
           and L1.def = "Action OR result of the action of L1.value")
           then L2.def = "Perform the L2.type"
```

**Cross-definitions.** Cross-definitions for indirectly related  $(L_1, L_2)$  and  $(L_2, L_1)$  pairs are generated in the same manner. We use definition patterns that have been manually specified in accordance with  $L_1$  and  $L_2$  morphological structures and parts of speech. For instance, for a  $(Xage_N, Xeur_N)$  pair, the schema for the  $Xage_N$  definition is “Action performed by  $Xeur_N$ .<sup>7</sup>” The definitions thus created are then assigned

---

<sup>5</sup>The situation is different for  $L_1$  nouns ending in *-eur* that are related to an adjectival  $L_2$  (e.g., *pâleur* ‘paleness’  $\leftarrow$  *pâle* ‘pale’). Here, the correct semantic type for the noun is “property.” Hence, both the  $L_1$  suffix and the  $L_2$  part of speech are required values to prevent over-generation.

by means of either target values (21a) or semantic types (21b) to the  $(L_1, L_2)$  pairs that satisfy the specified conditions of part of speech and word structure:

- (21) a. if  $(L1.\text{cat} = \text{"NOM"} \text{ and } L1.\text{structure} = \text{"Xage"})$   
 $\quad \quad \text{and } L2.\text{cat} = \text{"NOM"} \text{ and } L2.\text{structure} = \text{"Xeur"})$   
 $\quad \text{then } L1.\text{def} = \text{"Action performed by } L2.\text{value"}$
- b. if  $(L1.\text{cat} = \text{"NOM"} \text{ and } L1.\text{structure} = \text{"Xage"})$   
 $\quad \quad \text{and } L2.\text{cat} = \text{"NOM"} \text{ and } L2.\text{structure} = \text{"Xeur"})$   
 $\quad \text{then } L1.\text{def} = \text{"Action performed by } L2.\text{type"}$

## 1.4 Démonette: major semantic features

We now know how the Démonette architecture is constructed and how its basic information is obtained and can focus on the linguistic underpinnings of its morphological and semantic features.

### 1.4.1 Cumulative semantics

In the current version of Démonette, direct relations between derivatives and their bases are labeled with concrete definitions obtained from DériF. An opposite definition is then obtained by reversing this definition, as explained below (Section 1.4.3). Furthermore, the meanings ascribed to most indirect relations can be induced from these definitions, inasmuch as these relations are semantically and pragmatically relevant (see further discussion on this topic in Section 1.6).

Here, we assume that a morphological meaning is an aggregation of redundant elementary meanings, where an elementary meaning is the semantic contribution of a single morphological relation (e.g., the processive meaning of a deverbal noun such as *production*). Therefore, each morphological relation contributes to the meanings of the words it connects. Elementary meanings then combine to produce overall meanings. Redundancy may originate from the composition of elementary relations, as in (22), where the relation (and thus definition) in (22c) is deduced from relations (22a) and (22b).

- (22) a. *momifiable/ADJ*  $\leftarrow$  *momifier/VER* : *que l'on peut momifier*  
‘mummifiable  $\leftarrow$  mummify: that which may be mummified’
- b. *momifier/VER*  $\leftarrow$  *momie/NOM* : *transformer en momie*  
‘mummify  $\leftarrow$  mummy: transform into a mummy’
- c. *momifiable/ADJ*  $\leftarrow$  *momie/NOM*: *que l'on peut transformer en momie* ‘mummifiable  $\leftarrow$  mummy: that which may be transformed into a mummy’

Similarly, the meaning of a complex word derived from the same base as or having a common morphological ancestor with another word may

generally be expressed with respect to the meaning of the latter. For instance, *momifiable* can be defined relative to *momifier* (22a), and *momification* is also connected to *momifier*, which can be paraphrased as in (23a). It follows that *momifiable* and *momification* are cross-defined in relation to each other (23b, 23c):

- (23) a. *momification/NOM* ← *momifier/VER* : *action de momifier*  
                  ‘mummification ← mummify: action of mummifying’
- b. *momifiable/ADJ* ← *momification/NOM* : à qui il est possible  
                  d’appliquer la *momification* ‘mummifiable ← mummification:  
                  to which it is possible to apply mummification’
- c. *momification/NOM* ← *momifiable/ADJ* : *acte applicable à*  
                  *ce(lui) qui est momifiable* ‘mummification ← mummifiable:  
                  action applicable to that which is mummifiable’

As seen in this example, we use morphology to interconnect the elementary meanings of related complex words through shared predicates and arguments. This ability is implemented in Démonette such that each word receives as many definitions as it has connections with members of its morphological family.

#### 1.4.2 Derivational series and semantic types

Morphological relations and the lexical units they connect can be grouped into semantic classes (or types). For instance, *-ion* (*manifestation* ‘demonstration’), *-age* (*lavage* ‘washing’) and *-ment* (*soulèvement* ‘raising’) derivatives belong to the semantic type of action nouns. With this level of abstraction, we have a method of comparing lexical units and their meanings. For instance, the relations *manifester* ‘demonstrate’ → *manifestation*; *laver* ‘wash’ → *lavage*; and *soulever* ‘raise’ → *soulèvement* share the same abstract definitions and connect lexemes that belong to the same semantic classes. A definition expresses the meaning of a relation target with respect to the source of the relation. In an abstract definition, the meaning of the source is represented by a typed variable that is identified by its semantic type. Two types of identifiers are used for these semantic types:

- the semantic type of a (verbal) predicate such as *danser* ‘dance’ is denoted by @.
- the semantic type of a (derived) noun or adjective is denoted by an @ followed by an identifier of the role the derivative plays with respect to the predicate. For example, the suffixed noun *danseur<sub>(M)</sub>*, which represents the (M)asculine (AG)ent of the action predicate *danser*, is labeled @AGM. Likewise, the feminine noun *réparatrice* ‘repairer<sub>(F)</sub>’ is labeled @AGF, meaning (F)eminine (AG)ent, and

the (ACT)ion noun *réparation* ‘repairing’ is labeled @ACT. We use distinct labels for verbal and deverbal processive nouns because they differ in some of their semantic properties (Pustejovsky, 1995).

The relations implemented in the current version of Démonette connect words from a set that includes the derivational series of the suffixes listed in Table 1 and their bases. These words are verb predicates that denote actions, nouns that denote the same actions or agents that are capable of realizing them, and adjectives that denote properties. In their corresponding derivational families, the semantic representations of the nouns and adjectives involve the verb predicate *p*. The situation to which *p* refers can be regarded as an action or a state (see Table 2).

Tag	Meaning	Series	Examples
@ACT	ACTion noun	<i>Xage</i> <i>Xment</i> <i>Xion</i>	<i>lavage</i> ‘washing’ <i>gonflement</i> ‘inflation’ <i>exclusion</i> ‘exclusion’
@AGM	Masculine AGent noun	<i>Xeur</i>	<i>danseur</i> ‘dancer <sub>(M)</sub> ’ <i>sauveteur</i> ‘rescuer <sub>(M)</sub> ’
@AGF	Feminine AGent noun	<i>Xeuse</i> <i>Xrice</i>	<i>danseuse</i> ‘dancer <sub>(F)</sub> ’ <i>sculptrice</i> ‘sculptor <sub>(F)</sub> ’
@PROP	PROPerty adjective	<i>Xif</i>	<i>combatif</i> ‘combative’
@	predicate	—	<i>combattre</i> ‘fightv’ <i>laver</i> ‘wash’ <i>danser</i> ‘danse’ <i>sculpter</i> ‘sculpt’ <i>gonfler</i> ‘swell’ <i>exclure</i> ‘exclude’

TABLE 2 Semantic types of Démonette lexical units.

**Action nouns.** French deverbal action nouns may be formed through suffixation with *-age*, *-ment*, *-ion*, *-ure* or *-is* (Dal et al., 2004, Fradin, 2011, Haas et al., 2008, Kerleroux, 2008, Lecomte, 1997) or through conversion (as in *voler* ‘fly<sub>V</sub>’ → *vol* ‘flight<sub>N</sub>’ (Tribout, 2010, 2012) or *entrer* ‘enter’ → *entrée* ‘entry’ (Ferret and Villoing, 2012)). We chose to first focus on those derivations that are

- most easily identifiable (which excludes conversion, the orientation of which is often indeterminate; see Tribout (2010)),
- most available (which, for example, excludes the verb-to-noun *-is* suffixation, as in *semerv* ‘sow’ → *semis<sub>N</sub>* ‘seed,’ because it is no longer productive in synchrony according to the definition of productivity provided in Baayen (1992), Dal et al. (2008)), and

- most regular with respect to the derivative's semantic type (which excludes *-ure* because nouns derived by this suffix are often polysemous and may denote concrete entities; compare, e.g., *déchirer* 'rip<sub>V</sub>, tear<sub>V</sub>' → *déchirure* 'rip<sub>N</sub>', where the derived noun refers to the verb's action or its result, with *couvrir* 'cover<sub>V</sub>' → *couverture* 'coverage<sub>N</sub>' or 'blanket<sub>N</sub>', where the noun may also refer to a concrete entity).

For all these reasons, we selected the action nouns derived through suffixation with *-age*, *-ment* and *-ion*.

**Agent nouns.** Agent nouns are constructed via suffixation with *-eur*, *-euse* or *-rice*. The *-euse/-rice* distribution depends on the formal properties of the verb stem selected by the WFR: *-euse* is preferred for stems used for the imperfect tense, whereas *-rice* tends to select so-called learned roots, which are verb stems inherited from the supine stem of the Latin ancestor of the verb, or those that imitate this form (Bonami et al., 2009, Fradin and Kerleroux, 2003, Plénat, 2008, Roché, 2010).

**Property adjectives.** Among the derived adjectives that denote properties, we chose to include only those suffixed with *-if* and whose base is either a verb predicate (*combattre* 'fight<sub>V</sub>' → *combatif* 'combative') or a deverbal action noun related to a predicate, such as *abrasion* 'abrasion' → *abrasif* 'abrasive;' here, *abrasion* is itself a noun derived from the obsolete verbal form *abraser* 'abrade,' which is no longer attested in synchrony.

We also focused on these seven suffixes because these suffixes have been the subject of numerous descriptions; they are accounted for by both DériF and Morphonette (see Section 1.2), and they tend to construct derivatives that form partial derivational families, as in (24). Such a family is typically organized around a verb and contains indirect relations between family members that are located close to one another in the derivational graph; these indirect relations can be easily paraphrased, as in (25) for the relation between *prédateur* 'predator' and *prédatation* 'predation.'

- (24) a. *ravitailleur/VER, ravitaillement/NOM, ravitailleur/NOM, ravitailleuse/NOM*  
           'reupply, provision with supply, provider<sub>(M)</sub> of supply,  
           provider<sub>(F)</sub> of supply'
- b. *capter/VER, captage/NOM, captation/NOM, captatif/ADJ,*  
*captateur/NOM, capteur/NOM*  
           'harness<sub>V</sub>, channeling, harnessing, captivating, inveigler<sub>(M)</sub>,  
           sensor'
- c. *décantage/NOM, décantation/NOM, décantement/NOM, dé-*  
*canter/VER, décanteur/NOM*

- ‘decanting, decanting, decanting, decant, decanter’
- (25) *prédateur/NOM* : (*agent masculin habituel OR auteur masculin exceptionnel OR instrument*) de l’activité liée à la prédatation  
 ‘predator<sub>(M)</sub>: (usual masculine agent OR exceptional masculine agent OR instrument) of the action related to predation’
- (26) a. *prédateur/NOM, prédation/NOM, prédatrice/NOM*  
 ‘predator<sub>(M)</sub>, predation, predator<sub>(F)</sub>’  
 b. *audition/NOM, auditif/ADJ, auditeur/NOM, auditrice/NOM*  
 ‘audition, auditory, listener<sub>(M)</sub>, listener<sub>(F)</sub>’

These suffixes have another advantage: they also existed in Latin, from which they originate. This property allows Démonette to contain partial families, wherein the verb is lacking and the other members are inherited from Latin (26). We consider that Latinate lexemes that contain one of the suffixes listed in Table (1) belong to the derivational series corresponding to this suffixation rule, just as native French derivatives do. For example, the noun *audition* ‘audition’ is part of the series of nouns constructed with *-ion* in the same manner as, e.g., *application* ‘application’ or *préparation* ‘preparation.’ The inclusion of *audition* in this series is partially based on the fact that semantically, it denotes an action. It is also based on the formal properties of the noun, namely, the presence of the ending *-ion*, and on the fact that it is derivationally connected to a noun constructed with *-eur* (*auditeur* ‘listener<sub>(M)</sub>’), a noun constructed with *-rice* (*auditrice* ‘listener<sub>(F)</sub>’) and an adjective constructed with *-if* (*auditif* ‘auditory’), just as *application* is related to *applicateur* ‘applicator<sub>(M)</sub>,’ *applicatrice* ‘applicator<sub>(F)</sub>’ and *applicatif* ‘applicative,’ respectively. In addition, in families that contain nouns constructed with *-ion*, *-eur* and *-rice* and adjectives constructed with *-if*, the noun constructed with *-ion* always represents an action. In other words, although it can be etymologically identified as having been inherited from Latin, *audition* is a legitimate member of the derivational series of nouns constructed with *-ion*: it shares formal and semantic characteristics with such nouns, participates in the same derivational relations, and therefore shares the same semantic type. More generally, lexemes belonging to the same derivational series have the same semantic type.

The set of selected derivatives can be easily extended to adjectives constructed with *-oire*, such as *exploratoire* ‘exploratory;’ nouns constructed with *-oir*, such as *séchoir* ‘dryer;’ adjectives constructed with *-able*, such as *calculable* ‘calculable;’ and action or state nouns constructed with *-ance/-ence*, such as *partance* ‘departure’ or *souffrance*

‘pain.’ All of these derivatives are directly or potentially connected to a verb base.

- (27) a. *Cet enfant apprend à écrire avec beaucoup d’application*  
‘This child is learning to write with great diligence’
- b. *Nous allons procéder à une nouvelle application*  
‘We will produce a new application’

Theoretically, a lexeme can belong to several semantic-aspectual classes, which correspond to its different readings. For example, *application* has two meanings: ‘diligence’ and ‘application.’ It is both a property noun derived from the adjective *appliqué* ‘diligent’ (27a) and a noun derived from the action verb *appliquer* ‘apply’ (27b) (see Kerleroux (2008) for an analysis of stative nouns constructed with *-ion*). In the current version of Démonette, each lexeme has only one type because of the reduced number of affixations. In future work, the incorporation of a larger subset of TLFnome will entail accounting for multiple typing and the association of these types with different derivational relations and related definitions.

#### 1.4.3 Semantic annotations of direct and indirect relations

Relations between words are labeled with both concrete and abstract bi-oriented definitions. The direct concrete definitions were obtained from DériF’s analyses, whereas the others are specific to Démonette (see Figure 3). They are computed in accordance with the principles specified below. The semantic labels presented above are used as typed variables in the abstract definitions. Recall that abstract and concrete definitions characterize relations that connect either a derivative and its base (e.g., the direct relation *fonder* ‘found’ → *fondateur* ‘founder’) or two lexemes that share a common antecedent (i.e., indirect relations, such as *fondation* ‘foundation’ → *fondateur*).

**Principles.** Concrete, opposite and cross-definitions as well as all abstract definitions used to annotate direct and indirect relations obey the following principles:

1. Each word in Démonette is labeled with its **semantic type**. This type combines the semantic class to which the word belongs with the thematic role it plays with respect to the predicate to which it is related.
2. Each member of an indirect relation is defined with respect to the other member. An indirect relation includes two concrete **cross-definitions**, which are symmetrical to each other, and their corresponding abstract definitions.
3. There are four types of definitions on direct relations:

- (a) In a **concrete oriented definition**, the derivative is defined relative to the meaning of its base (see Section 1.2.2).
  - (b) In an **abstract oriented definition**, the base is represented by a typed variable.
  - (c) A **concrete opposite definition** is a paraphrasing of the base in relation to the meaning of the derivative. Therefore, a direct relation has a bi-oriented definition just as indirect relations do.
  - (d) An **abstract opposite definition** is an abstract counterpart of a concrete opposite definition.
4. Abstract oriented, opposite or cross-definitions of a pair of words ( $X, Y$ ) involve not  $X$  and  $Y$  but their **semantic labels**, thus achieving an abstraction of the relation between  $X$  and  $Y$ .

**Oriented, opposite and cross-definitions.** An oriented (resp., opposite) definition expresses the semantic link that directly connects a derivative to its base (resp., vice-versa), whereas a cross-definition expresses the link that connects it to another member of its morphological family. In what follows, we demonstrate the design and implementation of abstract definitions, either from DériF’s analyses or directly within Démonette. Abstract definitions are obtained from concrete definitions through the simple replacement of the lexical values of the derivational family members with semantically typed variables. In practice, an abstract definition contains only one typed variable to represent the other member of the underlying morphological relation.

For the derivational families considered in Démonette, the implementation of the abovementioned principles is illustrated in Table 3; abstract definitions are provided in column 2 and always involve the @ parameter. Disjunctive definitions, those which contain “OR,” currently cover the following situations:

- Deverbal nouns constructed with *-eur* may denote a usual masculine agent (*un marcheur* ‘a walker<sub>(M)</sub>’), an exceptional masculine agent (*le vainqueur* ‘the winner<sub>(M)</sub>’) or an instrument (*le dériveur* ‘the dinghy’) of the base predicate.
- The same holds for the *-euse* and *-rice* variants of the WFRs that form feminine agent nouns and instruments.
- Nouns constructed with *-age*, *-ion* and *-ment* may refer to an action (*lavage*) or the result of this action (*jaunissement* ‘yellowing’).

Column 1 in Table 3 provides the schema of the direct derivative/base relation, column 2 provides the abstract oriented definition of the derivative with respect to its base, and column 3 provides the

Oriented definition	Opposite definition	Example
$X_{eur_N}:X_V$		
(agent masculin habituel OR auteur masculin exceptionnel OR instrument) de @ (usual masculine agent OR exceptional masculine agent OR instrument) of @'	réaliser l'activité dont l'agent masculin est le @AGM 'perform the activity of which the masculine agent is @AGM'	marcheur:marcher 'walker <sub>(M)</sub> :walk'
$(X_{euse_N} X_{rice_N}):X_V$		
(agent féminin habituel OR auteur féminin exceptionnel OR instrument) de @ (usual feminine agent OR exceptional feminine agent OR instrument) of @'	réaliser l'activité dont l'agent féminin est le @AGF 'perform the activity of which the feminine agent is @AGF'	sculptrice: sculpter 'sculptor <sub>(F)</sub> :sculpt'
$X_{age_N}:X_V \ X_{ion_N}:X_V \ X_{ment_N}:X_V$		
(action OR résultat de l'action) de @ (action OR result of the action) of @'	réaliser le/la @ACT 'perform the @ACT'	abaissement: abaisser 'lowering:lower'
$X_{if_A}:X_V$		
en rapport avec l'acte de @ 'in connection with the act of @'	manifester le fait d'être @PROP 'manifest being @PROP'	combatif:combattre 'combative:combat'

TABLE 3 Oriented and opposite abstract definitions implemented in Démonette.

abstract opposite definition of the base with respect to its derivative. The opposite definitions are expressed symmetrically with respect to the oriented ones.

Cross-definitions participate in the combination of the elementary meanings introduced above. A complex lexeme can be defined with respect to another complex lexeme with which it has an indirect relation when this relation involves a reasonably low number of intermediaries (regarding this point, see Section 1.6). In the derivational families of the current version of Démonette, all indirect relations are relevant and annotated with a cross-definition (see Table 4). Indirect relations may connect lexemes belonging to the same semantic type (excluding

gender). Three types of relations share this property:

1. the connected lexemes may be doublets formed via concurrent WFRs (see Plag (1999) on the typology of suffix rivalry), e.g., *lavage* ↔ *lavement* ‘enema;’ we consider these doublets to be synonyms in the broad sense (see, e.g., Cruse (2004) on synonymy classification);
2. the indirect relations may connect derived words formed from the same base with the same suffix using two distinct stem variants of the base, e.g., *ausculteur* ‘auscultator’<sub>(M)</sub> ↔ *auscultateur* ‘auscultator’<sub>(M)</sub>; we assume, as above, that these doublets are synonymous in the broadest sense (but see Anscombe (2003) on the difference in meaning between *-eur* and *-ateur* nouns derived from the same verb); and
3. the connected lexemes may each constitute a pole in the binary male ↔ female opposition, when one is constructed with *-eur* and the other with *-euse* or *-rice*.

Cross-definition <sub>1:2</sub>	Cross-definition <sub>2:1</sub>	Example
(Xage Xion Xment):Xeur		
<i>action</i> <i>pratiquée</i> <i>par</i> @AGM 'action'    performed    by @AGM'	<i>(agent masculin OR instrument) du @ACT</i> (masculine agent OR instrument) of @ACT	<i>distraction:</i> <i>distracteur</i> 'distraction:' <i>distractor<sub>(M)</sub></i> '
(Xage Xion Xment):(Xeuse Xrice)		
<i>action</i> <i>pratiquée</i> <i>par</i> @AGF 'action'    performed    by @AGF'	<i>(agent féminin OR instrument) du @ACT</i> (feminine agent OR instrument) of @ACT	<i>fondation:</i> <i>fondatrice</i> 'foundation:' <i>founder<sub>(F)</sub></i> '
Xif:(Xage Xion Xment)		
<i>qui permet la @ACT</i> 'enabling the performance of @ACT'	<i>action de ce qui est @PROP</i> 'action of one who/that is @PROP'	<i>déterminatif:</i> <i>détermination</i> 'determinative:' <i>determination</i> '
Xif:Xeur		
<i>qui caractérise l'activité pratiquée par @AGM</i> 'which characterizes the activity performed by @AGM'	<i>celui dont l'activité est @PROP</i> 'one <sub>(M)</sub> whose activity is @PROP'	<i>administratif:</i> <i>administrateur</i> 'administrative:' <i>administrator<sub>(M)</sub></i> '
Xif:(Xeuse Xrice)		
<i>qui caractérise l'activité pratiquée par @AGF</i>	<i>celle dont l'activité est @PROP</i>	<i>spoliatif:</i> <i>spoliatrice</i>

Cross-definition <sub>1:2</sub>	Cross-definition <sub>2:1</sub>	Example
‘which characterizes the activity performed by @AGF’	‘one <sub>(F)</sub> whose activity is @PROP’	‘spoliative: spoliator <sub>(F)</sub> ’
<i>(Xeuse Xrice):Xeur</i>		
<i>celle qui a pour correspondant masculin @AGM</i> ‘which has @AGM as a masculine counterpart’	<i>celui qui a pour correspondant féminin @AGF</i> ‘which has @AGF as a feminine counterpart’	<i>danseuse:</i> <i>danseur</i>  ‘dancer <sub>(F)</sub> : dancer <sub>(M)</sub> ’
<i>(Xage:Xion Xage:Xment Xion:Xment)</i>		
@ACT	@ACT	<i>ruminement:</i> <i>rumination</i> ‘rumination: rumination’
<i>Xeur:X'eur</i>		
@AGM	@AGM	<i>activateur:</i> <i>activeur</i> ‘activator <sub>(M)</sub> : activator <sub>(M)</sub> ’
<i>Xrice:X'euse</i>		
@AGF	@AGF	<i>activatrice:</i> <i>activeuse</i> ‘activator <sub>(F)</sub> : activator <sub>(F)</sub> ’

TABLE 4: Cross-definitions implemented in Démonette.

In the last two rows of Table 4,  $X'$  and  $X$  are two stem variants of the same base. Column 1 provides the schema of the indirect relation derivative<sub>1</sub>:derivative<sub>2</sub>, column 2 provides the definition of derivative<sub>1</sub> with respect to derivative<sub>2</sub>, and column 3 provides the definition of derivative<sub>2</sub> with respect to derivative<sub>1</sub>. The table presents only one of the indirect relations. The other can be obtained by reversing the two members of the schema (e.g.,  $Xeur:(Xage|Xion|Xment)$  for the first group header) and switching columns 1 and 2.

## 1.5 Démonette’s content

In its current version, Démonette contains 31,204 entries. These entries all belong to the derivational series constructed using the seven suffixes listed in Table 1. DériF has provided 21,556 pairs of family members through the matching of the antecedents of lexemes constructed on a

verb predicate. For example, the antecedents computed in the DériF analyses of *lavement* (28a), *lavage* (28b), *laveur* ‘wisher’<sub>(M)</sub> (28c) and *laveuse* ‘wisher’<sub>(F)</sub> (28d) indicate that all these words share the verb *laver* as common predicate and can be combined into the family (29). Then, a total of 30 pairs of words can be extracted from this family, such as those in (30).

- (28) a. (*lavement/NOM, laver/VER*)
- b. (*lavage/NOM, laver/VER*)
- c. (*laveur/NOM, laver/VER*)
- d. (*laveuse/NOM, laver/VER*)
- (29) / *lavage/NOM=@ACT, lavement/NOM=@ACT, laver/VER=@, laveur/NOM=@AGM, laveuse/NOM=@AGF* /
- (30) a. *lavage/NOM=@ACT laver/VER=@*  
          @*ACT = (action OR résultat de l'action) de @*  
          @ = *réaliser le/la @ACT*
- b. *lavage/NOM=@ACT lavement/NOM=@ACT*  
          @*ACT = @ACT*
- c. *lavage/NOM=@ACT laveur/NOM=@AGM*  
          @*ACT = action pratiquée par @AGM*  
          @*AGM = agent masculin OR instrument du @ACT*

The additional contribution of Morphonette consists of 9,648 relations, 1,145 of which are direct, with the remainder being indirect. The corresponding cross-definitions were added to Démonette following the same principles described above.

Each Démonette entry describes a pair of morphologically related words. Each entry has 13 fields, of which only 9 are described below. In what follows, concrete definitions are omitted.

- $w_1$  and  $w_2$  are the two member words of the pair;
- O(igin) is the source of the relation (D for DériF and M for Morphonette);
- ST contains the semantic types of  $w_1$  and  $w_2$ ;
- the values of “Cross Def” are the cross-definitions of  $w_1$  relative to the meaning of  $w_2$  and, conversely, of  $w_2$  relative to the meaning of  $w_1$ ;
- if  $w_1$  and  $w_2$  are indirectly related, “Oriented Def” describes  $w_1$  and  $w_2$  with respect to their common antecedent, whereas if one is derived from the other, “Oriented Def” contains the appropriate oriented and opposite definitions.

In the latter case, the field labeled “Cross Def” is empty. Tables 5 and 6 show the entries for the pairs *administratif:administratrice* ‘administrative:administrator<sub>(F)</sub>’ and *administratrice:administrer* ‘administrator<sub>(F)</sub>:administrate.’

$w_1 / w_2$	O	ST	Oriented Def	Cross Def
<i>administratif<sub>A</sub></i>	M	@PROP	En rapport avec l'acte de @	Qui caractérise l'activité de @AGF
<i>administratrice<sub>N</sub></i>		@AGF	Agent féminin OR instrument de @	Celle dont l'activité est @PROP

TABLE 5 Description of the (*administratif, administratrice*) relations.

$w_1 / w_2$	O	ST	Oriented Def	Cross Def
<i>administratrice<sub>N</sub></i>	D	@AGF	Agent féminin OR instrument de @	
<i>administrer<sub>V</sub></i>		@	Réaliser l'action dont l'agent féminin est @AGF	

TABLE 6 Description of the (*administratrice, administrer*) relation.

At present, 8,376 Démonette relations originate jointly from DériF and Morphonette. The exclusive contribution of Morphonette constitutes 1,504 relations, all indirect. In other words, all the direct relations originating from Morphonette (i.e., between a verb and a derived noun or adjective) are also found in DériF. The 13,180 remaining relations were provided by DériF. The contributions originating from DériF include 8,802 direct relations and 4,378 indirect ones. Tables 7 and 8 summarize these figures. The examples presented in (31) and the de-

	pairs
DériF	21,556
Morphonette	9,648
Démonette	31,204

TABLE 7 Démonette’s content: the contributions of DériF and Morphonette  
tailed description of *abréacteur* ‘abreactor’ ↔ *abreaction* ‘abreaction’

relation type	$D \setminus M$	$D \cap M$	$M \setminus D$
direct	8,802	1,145	0
indirect	4,378	7,231	1,504

TABLE 8 Distributions of direct and indirect relations in Démonette with respect to their origins:  $D \setminus M$  is the number of pairs present only in DériF,  $M \setminus D$  is the number of pairs present only in Morphonette and  $D \cap M$  is the number of pairs present in both DériF and Morphonette.

provided in Table 9 illustrate the sort of word pairs found in Démonette that were provided by Morphonette and are absent from DériF.

- (31) a. *acteur<sub>N</sub>:actrice<sub>N</sub>* ‘actor:actress’
- b. *abréacteur<sub>N</sub>:abréaction<sub>N</sub>* ‘abreactor:abreaction’
- c. *abréviateur<sub>N</sub>:abréviatif<sub>A</sub>* ‘abbreviator:abbreviative’
- d. *admonition<sub>N</sub>:admonitrice<sub>N</sub>* ‘admonition:admonisher<sub>(F)</sub>’

$w_1 / w_2$	O	ST	Oriented Def	Cross Def
<i>abréacteur<sub>N</sub></i>	M	@AGM		Agent féminin OR instrument de @
<i>abréaction<sub>N</sub></i>		@ACT		Action pratiquée par @AGM

TABLE 9 Description of the (*abréacteur*, *abréaction*) relation.

Conversely, there are at least two explanations for the absence from Morphonette of 4,378 indirect relations acquired from DériF. The first is parsing errors. These may be attributed to the behavior of DériF, which, for example, wrongly considers *compassement* ‘measure with a compass’ and *compassion* ‘compassion’ to belong to the same morphological family. In fact, *compassement* is derived from *compas* ‘compass,’ whereas *compassion* is a deverbal noun derived from *compatir* ‘sympathize.’ These errors occurred during the creation of the indirect relations. Similarly, *accoutreur* ‘costumer<sub>(M)</sub>’ and *raccourage* ‘mending’ should not form an entry in Démonette because the second word contains a prefix *r-* that is absent in the first.

The second explanation is the differences between the approaches implemented by the two systems. Indeed, Morphonette uses a threshold related to the number of pairs in an analogical series that excludes less frequent combinations. Thus, as shown in (32), the absence from Morphonette of some of the pairs obtained from DériF is attributable to the low frequency of their formal pattern.

- (32) a. *abaissement/NOM:abaisseur/NOM* ‘lowering:one who lowers’  
       b. *abattage/NOM:abattement/NOM* ‘slaughter:reduction’  
       c. *abolissement/NOM:abolitif/ADJ* ‘abolishment:abolishing’  
       d. *abortion/NOM:avorteur/NOM* ‘abortion:abortionist’

<i>Xteur</i>	158						
<i>Xeuse</i>	<b>3,938</b>	134					
<i>Xrice</i>	338	<b>978</b>	66				
<i>Xment</i>	<b>2,226</b>	98	<b>1,390</b>	40			
<i>Xage</i>	<b>2,946</b>	80	<b>1,932</b>	11	<b>1,656</b>		
<i>Xion</i>	<b>926</b>	<b>1,544</b>	208	<b>976</b>	320	240	
<i>Xif</i>	310	448	36	304	26	24	<b>918</b>
	<i>Xeur</i>	<i>Xteur</i>	<i>Xeuse</i>	<i>Xrice</i>	<i>Xment</i>	<i>Xage</i>	<i>Xion</i>

TABLE 10 Distribution of indirect relations with respect to suffixes.  
       Numbers greater than 900 are printed in bold.

Table 10 summarizes the distribution of indirect relations in Démonette with respect to the suffixes involved. Nouns constructed with *-eur* are labeled as *Xteur* (e.g., *agitateur* ‘stirrer’) when formed using *X*’s learned root, cf. Section 1.4.2. This notation is used to distinguish them from the nouns constructed with *-eur* that are labeled as *Xeur*, for which the verb imperfect stem is used (e.g., *régisseur* ‘manager’). Numerical disparity is evident in the occurrence of these relations.

(a) The most commonly represented indirect relations reflect a particular preference for some suffixes to be used with the same types of stems: the suffixes *-ion*, *-if* and *-rice* are used with learned roots (and are therefore called “learned suffixes”), whereas the suffixes *-ment* and *-age* tend to select ordinary or native stems. More precisely, they are combined with the verb stem that is used to form the imperfect tense, cf. Bonami et al. (2009). As shown in the first two columns of Table 10, *-eur* exhibits both behaviors. In this regard, we note that synonymous pairs referring to masculine agents (158 *Xeur:Xteur* pairs) are almost three times as numerous as those denoting feminine agents (66 *Xeuse:Xrice* pairs). This discrepancy can be explained by the smaller number of feminine agent nouns recorded in TLFnome (and also in the Démonette word list) with respect to the corresponding masculine nouns: indeed, there are 13,388 relations involving a noun constructed with *-eur* and 5,444 involving a noun constructed with *-teur*, but only 9,820 involving a noun constructed with *-euse* and 3,226 involving a noun constructed with *-rice*.

- (b) Some results vary within a range of 200 to 900 word pairs: they are often  $X\text{suff}_1:X\text{suff}_2$  pairs, where  $\text{suff}_1$  or  $\text{suff}_2$  is a learned suffix whereas the other is not:  $X\text{eur}:X\text{rice}$ ,  $X\text{eur}:X\text{if}$ ,  $X\text{teur}:X\text{euse}$ ,  $X\text{euse}:X\text{ion}$ ,  $X\text{ment}:X\text{ion}$  and  $X\text{age}:X\text{ion}$ . The relatively small number of  $X\text{teur}:X\text{if}$  and  $X\text{rice}:X\text{if}$  learned pairs, those that involve an agent noun (be it masculine or feminine) and a property adjective, is a consequence of the fact that adjectives constructed with *-if* are underrepresented in Démonette with respect to nouns constructed with *-ion* (1,288 for the first compared with 8,484 for the second). This is partially attributable to the competition between *-if* and *-oire* (e.g., *vexatoire* ‘vexatious’) in the formation of learned property adjectives whose bases refer to actions.
- (c) Those patterns of ‘mixed’ suffix pairs (i.e., one learned suffix and one ordinary one) that are composed of a learned masculine agent ( $X\text{teur}$ ) and a native action noun, i.e.,  $X\text{ment}$  or  $X\text{age}$ , constitute a few hundred indirect relations; in general, mixed pairs that consist of a masculine agent noun and its related action noun contain almost exclusively (native)  $X\text{eur}$  and (learned)  $X\text{ion}$  nouns. Finally, some relationships are particularly infrequent:  $X\text{rice}:X\text{ment}$ ,  $X\text{rice}:X\text{age}$ ,  $X\text{ment}:X\text{if}$ , and  $X\text{age}:X\text{if}$ . Several factors that have already been discussed serve to explain these low figures: (i)  $X\text{rice}$  is less represented than  $X\text{teur}$ , (ii) *-if* probably suffers from its rivalry with *-oire*, and (iii) mixed pairs that involve an agent noun and an action noun favor *-ion* for the latter.

These results demonstrate how formal combinations can be ranked based on their frequency, which provides some idea of which suffix pairs are most commonly used to produce morphological indirect relations between masculine and feminine agent nouns, agent and action nouns, or an agent noun and a property adjective. Of course, these hypotheses must be validated through projection onto a large-scale text corpus.

We conducted a partial assessment of Démonette by comparing its content to that of Verbaction (Tanguy and Hathout, 2002). Verbaction is a database that consists of 9,386 verb/action noun pairs whose members are morphologically related; its content has been entirely validated manually. The experiment, reported in (Hathout and Namer, 2014b), involved a subset of Démonette that includes only action nouns and their verbal bases. With respect to this subset, we obtained 84% recall and 90% precision.

## 1.6 Discussion

The current version of Démonette only partially spans the French lexicon and includes only a small percentage of DériF and Morphonette

analyses. In future development of Démonette, we will gradually increase its coverage to incorporate all of these analyses. This expansion will involve new developments and challenges that were not addressed in the current version. Subsequently, we will also integrate all TLFnome entries and then entries from other large French lexicons, such as Lefff (Sagot et al., 2006) and GLÀFF (Sajous et al., 2013, Hathout et al., 2014).

On the one hand, we must increase the number of relevant lexical semantic types. The identification of additional types will benefit from the extensive literature that exists concerning thematic roles (Fillmore, 1968) and qualia (Pustejovsky, 1995) as well as actantial (Foley and Van Valin, 1984) and lexical relations (Cruse, 1986) and lexical functions (Mel'čuk, 1996). The computation of oriented, opposite and cross-definitions as well as concrete and abstract definitions will be straightforward, as DériF already provides the former. The implementation of polysemous lexemes such as *application* (see Section 1.2.2) will benefit from the high redundancy of their representations. Nothing prevents a lexeme from having several oriented and/or opposite definitions or several semantic types.

The generation of indirect relationships within derivational families and, above all, of cross-definitions remains a difficult issue. From a formal point of view, the problem is not difficult to solve: it should be possible to form derivational families into complete graphs in which all lexemes are interconnected. However, several questions arise: will the contributions of all indirect relations be equally meaningful to the lexicon? Are all of them worthwhile, as far as psychological plausibility is concerned, i.e., how do speakers use them to organize their derivational lexicon? Is relevance a matter of distance in the derivational graph between words of the same family only? For example, in the derivational family example presented in (33), the definition of the edge between *activateur* and *activation* is justified (especially because it is spontaneously recovered by speakers, who regard *activateur* as the noun referring to the author of an *activation* and, conversely, regard the noun *activation* as the act performed by the *activateur*), whereas defining the noun *activateur* with respect to the adverb *activement* seems less legitimate.

- (33) (*actif<sub>A</sub>*, *activer<sub>V</sub>*, *activateur<sub>N</sub>*, *activatrice<sub>N</sub>*, *activation<sub>N</sub>*, *activiste<sub>A</sub>*, *activisme<sub>N</sub>*, *activement<sub>Adv</sub>*, etc.)  
 '(active, activate, activator<sub>(M)</sub>, activator<sub>(F)</sub>, activation, activist, activism, actively, etc.)'

In the absence of any known studies of these indirect relations, we

propose that their coverage should be limited using several criteria, such as the distance between connected vertices in the graph of direct relations. This distance depends on the pair under consideration. By default, it should be 2, but it may be 3 for derived words that contain very frequent affix sequences, e.g., verbs ending with *-aliser*, such as *populariser* ‘popularize’ (Namer, 2013a, Hathout and Namer, 2014a); nouns constructed with *-isation*, such as *nationalisation* ‘nationalization;’ and adjectives prefixed with *in-* and suffixed with *-able*, such as *incontournable* ‘inescapable’ (Dal and Namer, to appear).

A second criterion, which is related to the previous one, is to consider that a “useful” indirect relation is a relation for which speakers can readily formulate a cross-definition. Our assumption, which still needs to be verified experimentally, is that the indirect relations considered in the current version of Démonette satisfy this condition (see Section 1.5). However, other indirect relations are more difficult to define in a sufficiently regular manner for integration into Démonette. For example, this is the case for pairs of masculine and feminine nouns constructed with *-et:-ette* or *-ier:-ière*, such as *cachet:cachette* or *boulevardier:boulevardière*. A *cachette* ‘hiding place’ is not a feminine version of *cachet* ‘seal,’ and a *boulevardière* ‘prostitute’ is not a female *boulevardier* ‘author of bedroom farces.’ The criterion related to the distance in the graph does not always accurately predict the possibility of creating cross-definitions.

Other pairs are difficult to define relative to one another, such as

- adverbs constructed with *-ment* and verbs constructed with *-iser* that are formed from the same adjectival base, such as *stérile* ‘sterile,’ *total* ‘total’ and *verbal* ‘verbal’ (34), and
- deverbal adjectives constructed with *-able* with bases of the form *Xiser*, where *X* is an adjective, and nouns constructed with *-iste* that are formed from the same *X*, such as *canonique* ‘canonical,’ *conceptuel* ‘conceptual’ and *individuel* ‘individual’ (35).

(34)	<i>stérile<sub>A</sub></i>	→	<i>stérilement<sub>Adv</sub>:stériliserv<sub>V</sub></i>
	‘sterile’		‘fruitlessly:sterilize’
	<i>total<sub>A</sub></i>	→	<i>totalement<sub>Adv</sub>:totaliserv<sub>V</sub></i>
	‘total <sub>A</sub> ’		‘totally:total <sub>V</sub> ’
	<i>verbal<sub>A</sub></i>	→	<i>verbalement<sub>Adv</sub>:verbaliserv<sub>V</sub></i>
	‘verbal’		‘verbally:verbalize’

- (35) *canonique<sub>A</sub>* → *canonisable<sub>A</sub>:canoniste<sub>N</sub>*  
     ‘canonical’                         ‘canonizable:canonist’  
*conceptuel<sub>A</sub>* → *conceptualisable<sub>A</sub>:conceptualiste<sub>N</sub>*  
     ‘conceptual’                         ‘conceptualizable:conceptualist’  
*individuel<sub>A</sub>* → *individualisable<sub>A</sub>:individualiste<sub>N</sub>*  
     ‘individual’                         ‘individualizable:individualist’

Another difficulty that we have not yet addressed is the integration of lexemes constructed through conversion or compounding. Both morphological constructions are analyzed by DériF. The inclusion in Démonette of oriented and opposite definitions for converted words will be quite straightforward because these definitions follow the same patterns as affixed words. Conversely, the integration into Démonette of definitions of compounds may be more difficult to perform because relations involved by compound words are not binary but ternary. An additional difficulty arises from the fact that compounding elements may be bound roots, such as *hydro* ‘hydro = water’ in *hydromassage* ‘hydromassage.’ Their representation in Démonette will most likely need additional mechanisms.

Nevertheless, Démonette’s architecture is sufficiently flexible to allow for the resolution of the difficulties identified above. This flexibility also allows for facile integration of more specialized lexicons, such as Verbaction, which describes action nouns; the Lexeur lexicon, which contains agent nouns constructed with *-eur* (Fabre et al., 2004); or some of the features of nouns suffixed with *-aie*, *-at*, *-iste* and *-isme* from the *Dictionnaire des mots construits* (Roché, 2011).

The applications of Démonette in NLP are comparable to those of similar lexical resources that provide inflectional information, such as Morphalou (Romary et al., 2004); those that provide phonological descriptions, such as Lexique (New et al., 2004); those that provide syntactic annotations, such as Lefff (Sagot et al., 2006); and those that provide semantic annotations, such as EurowordNet (Vossen, 1998) or Wolf (Sagot and Fišer, 2008). The unique advantage offered by Démonette lies in the variety of morphological relations and morpho-semantic annotations recorded in the network. First, words are related to each other through multiple relations: a base word is connected to many derivatives, and derived words from the same family and series are related to each other. Second, each word is semantically typed, and its meaning is paraphrased via several definitions, one for each of the word’s relations with other words. Third, these features are also available for morphologically simple words, as all relations are bi-oriented. These three properties can be exploited for various tasks in language

technology and NLP applications. As noted in Clark et al. (2008), the analysis of textual content can be improved through the enrichment of lexical resources with semantic annotations. Furthermore, word-to-word relations and their associated definitions facilitate word selection and the identification of lexical meaning. Similar to WordNet and Wolf, these capabilities can be advantageous for natural language processing applications (e.g., word-sense disambiguation, information retrieval). Moreover, they can be used in tools designed for content production: the ability to choose between words based on the relations between them, their semantic types and their bases and to replace them with paraphrases is of significant benefit to computer-aided translation, text summarization, standardization and generation. In association with other distributional bases (Turney and Pantel, 2010), Démonette can also be used for lexical substitution tasks (McCarthy et al., 2004) and semantic categorization (Tsatsaronis and Panagiotopoulou, 2009).

## Conclusion

In this paper, we have presented the first version of Démonette, a lexical resource that currently contains 31,204 morphologically annotated entries and was constructed automatically from the results produced by two systems that implement two different approaches to Word Formation theories: Morphonette, which is based on formal analogies between lexemes and implements theoretical principles of relational and paradigmatic morphology, and DériF, which exploits oriented linguistic rules that have been manually designed and validated by linguists.

Words in Démonette that belong to the same derivational family are connected to each other through direct and indirect relations depending on their degree of relatedness. Relevant relations are annotated with bi-oriented (concrete) definitions that express the meaning of one lexeme relative to the meaning of another lexeme. Thus, a lexeme can potentially be described by as many definitions as it has relations with other lexemes in the lexicon. The lexemes are typed, and these types are used as parameters to associate abstract definitions with concrete definitions. Words that share the same abstract definition form derivational series. Abstract definitions can assist in identifying the contributions of each semantic type in the development of the Démonette morphological network.

Currently, Démonette contains only seven classes of derivatives and their corresponding verb bases. The constituent data has been drawn exclusively from dictionaries. Nevertheless, we believe that the methodology presented in this paper can be extended to data acquired from

corpora: DériF and Morphonette are designed to perform morphological (and semantics) analyses of unknown words. However, the design of new definitions, which will be used to label indirect relations between lexemes belonging to other derivational series, should account for the speaker's point of view, i.e., its relevance; in other words, we assume that definitions are useless if speakers cannot express them spontaneously. This usually occurs when definitions are too complex.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions for improving the quality of the paper.

## References

- Amiot, Dany. 2008. La catégorie de la base dans la préfixation en *dé-*. In B. Fradin, ed., *La raison morphologique. Hommage à la mémoire de Danielle Corbin*, vol. 27, pages 1–16. Amsterdam/Philadelphia: John Benjamins.
- Anscombe, Jean-Claude. 2003. L'agent ne fait pas le bonheur : agentivité et aspectualité dans certains noms d'agent en espagnol et en français. *Revista Complutense de Estudios Franceses* 11:11–27.
- Baayen, Harald. 1992. Quantitative aspects of morphological productivity. In G. E. Booij and J. van Marle, eds., *Yearbook of Morphology 1991*, pages 109–149. Dordrecht: Kluwer Academic Publishers.
- Baayen, R. Harald, Richard Piepenbrock, and 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 G. E. Booij and J. van Marle, eds., *Yearbook of Morphology 1993*, pages 1–25. Dordrecht: Kluwer Academic Publishers.
- Bernhard, Delphine. 2006. Automatic acquisition of semantic relationships from morphological relatedness. In T. Salakowski, F. Ginter, S. Pyysalo, and T. Pahikkala, eds., *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, vol. 4139 of *Lecture Notes in Computer Science*, pages 121–132. Berlin / Heidelberg: Springer Verlag.
- Bernhard, Delphine, Bruno Cartoni, and Delphine Tribout. 2011. A task-based evaluation of French morphological resources and tools. *Linguistic Issues in Language Technology* 5(2).
- Bilgin, Orhan, Özlem Çetinoğlu, and Kemal Oflazer. 2004. Morphosemantic relations in and across WordNets. In *Proceedings of the Global WordNet Conference*, pages 60–66.

- Bonami, Olivier, Gilles Boyé, and Françoise Kerleroux. 2009. L'allomorphie radicale et la relation flexion-construction. In B. Fradin, F. Kerleroux, and M. Plénat, eds., *Aperçus de morphologie du français*, pages 103–125. Saint-Denis: Presses universitaires de Vincennes.
- Clark, Peter, Christiane Fellbaum, Jerry R Hobbs, Phil Harrison, William R Murray, and John Thompson. 2008. Augmenting WordNet for deep understanding of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing (STEPS 2008)*, pages 45–57. Association for Computational Linguistics.
- Creutz, Mathias. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proceedings of the 41th annual meeting of the ACL*, pages 280–287. Sapporo, Japan: ACL.
- Creutz, Mathias and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Tech. Rep. A81, Helsinki University of Technology.
- Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Cruse, D. Alan. 2004. *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.
- Dal, Georgette, Bernard Fradin, Natalia Grabar, Fiammetta Namer, Stéphanie Lignon, and Pierre Zweigenbaum. 2008. Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats. In J. Durand, B. Habert, and B. Laks, eds., *1er Congrès Mondial de Linguistique Française*, pages 1587–1599. Paris: ILF.
- Dal, Georgette, Stéphanie Lignon, Fiammetta Namer, and Ludovic Tanguy. 2004. Toile contre dictionnaires : Analyse morphologique en corpus de noms déverbaux concurrents. In *Colloque international sur les noms déverbaux*. Villeneuve-d'Asq.
- Dal, Georgette and Fiammetta Namer. to appear. La fréquence en morphologie : pour quels usages ? *Langages* .
- De Pauw, Guy and Peter Waiganjo Wagacha. 2007. Bootstrapping morphological analysis of Gikuyū using maximum entropy learning. In *Proceedings of the Eighth Annual Conference of the International Speech Communication Association (Interspeech)*. Antwerp, Belgique.
- Fabre, Cécile, Franck Floricic, and Nabil Hathout. 2004. Collecte outillée pour l'analyse des emplois discordants des déverbaux en -eur. Talk given at the workshop *La place des méthodes quantitatives dans le travail du linguiste*. ERSS, Université de Toulouse II-Le Mirail.
- Fellbaum, Christiane, ed. 1999. *WordNet: an Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Fellbaum, Christiane and George A. Miller. 2003. Morphosemantic links in WordNet. *TAL. Traitement automatique des langues* 44(2):69–80.
- Fellbaum, Christiane, Anne Osherson, and Peter E. Clark. 2009. Putting semantics into WordNet's “morphosemantic” links. In *Human Language*

- Technology. *Challenges of the Information Society*, vol. 5603 of *Lecture Notes in Computer Science Volume*, pages 350–358. Springer.
- Ferret, Karen and Florence Villoing. 2012. L’aspect grammatical dans les nominalisations en français : les déverbaux en *-age* et *-ée*. *Lexique* 20:73–127.
- Fillmore, Charles. 1968. The case for case. In E. Bach and R. Harms, eds., *Universals in Linguistic Theory*, pages 1–88. New-York: Holt, Rinehart, and Winston.
- Foley, William A. and Robert D. J. Van Valin. 1984. *Functional Syntax and Universal Grammar*, vol. 38 of *Cambridge Studies in Linguistics*. Cambridge: Cambridge University Press.
- Fradin, Bernard. 2011. Remarks on state-denoting nominalizations. *Recherches linguistiques de Vincennes* 40:73–99.
- Fradin, Bernard and Françoise Kerleroux. 2003. Troubles with lexemes. In G. Booij, J. De Cesaris, S. Scalise, and A. Ralli, eds., *3d Mediterranean Morphology Meeting (MMM3) (selected papers)*, pages 177–196. Barcelona: IULA - Universitat Pompeu Fabra.
- Goldsmith, John. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics* 27(2):153–198.
- Gosme, Julien and Yves Lepage. 2009. A first study of the complete enumeration of all analogies contained in a text. In *4th Language & Technology Conference (LTC 2009)*, pages 401–405. Poznań, Poland.
- Haas, Pauline, Richard Huyghe, and Rafael Marín. 2008. Du verbe au nom : calques et décalages aspectuels. In J. Durand, B. Habert, and B. Laks, eds., *1er Congrès Mondial de Linguistique Française*, pages 2051–2065. Paris: ILF.
- Habash, Nizar and Bonnie Dorr. 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. ACL, Edmonton.
- Hammarström, Harald and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 32(2):309–350.
- Harris, Zellig. 1955. From phoneme to morpheme. *Language* 31(2):190–222.
- Hathout, Nabil. 2009. Acquisition of morphological families and derivational series from a machine readable dictionary. In F. Montermini, G. Boyé, and J. 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. In press. Phonotactics in morphological similarity metrics. *Language Sciences* [www.sciencedirect.com/science/article/pii/S0388000114000497](http://www.sciencedirect.com/science/article/pii/S0388000114000497).

- Hathout, Nabil and Fiammetta Namer. 2011. Règles et paradigmes en morphologie informatique lexématique. In *Actes de la 18e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2011)*. ATALA, Montpellier, France.
- Hathout, Nabil and Fiammetta Namer. 2014a. Discrepancy between form and meaning in word formation: the case of over- and under-marking in French. In F. Rainer, W. U. Dressler, F. Gardani, and H. C. Luschützky, eds., *Morphology and meaning (Selected papers from the 15th International Morphology Meeting, Vienna, February 2010)*, Current Issues in Linguistic Theory, pages 177–190. Amsterdam: John Benjamins.
- Hathout, Nabil and Fiammetta Namer. 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. ATALA, Marseille.
- Hathout, Nabil, Franck Sajous, and Basilio Calderone. 2014. GLÀFF, a large versatile French lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Kerleroux, Françoise. 2008. Des noms indistincts. In B. Fradin, ed., *La raison morphologique. Hommage à la mémoire de Danielle Corbin*, pages 113–132. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Lafourcade, Mathieu and Alain Joubert. 2013. Bénéfices et limites de l’acquisition lexicale dans l’expérience JeuxDeMots. In N. Gala and M. Zock, eds., *Ressources Lexicales: Contenu, construction, utilisation, évaluation*, vol. 30 of *Linguisticae Investigationes, Supplementa*, pages 187–216. Amsterdam: John Benjamins.
- Lecomte, Elsa. 1997. Tous les mots possibles en *-ure* existent-ils ? In D. Corbin, G. Dal, B. Fradin, B. Habert, F. Kerleroux, M. Plénat, and M. Roché, eds., *Mots possibles, mots existants*, vol. 1 of *Silexicales*, pages 191–200. Villeneuve d’Ascq: Presses de l’Université de Lille.
- Lepage, Yves. 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*, vol. 2, pages 728–735. Montréal.
- Lepage, Yves. 2004. Analogy and formal languages. *Electronic Notes in Theoretical Computer Science* 53:180–191. Proceedings of the the 6th Conference on Formal Grammar and the 7th on the Mathematics of Language (FG/MOL-2001).
- McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 280–287. Association for Computational Linguistics, Barcelona, Spain.
- Mel’čuk, Igor. 1996. Lexical functions: A tool for the description of lexical relations in the lexicon. In L. Wanner, ed., *Lexical Functions in*

- Lexicography and Natural Language Processing*, pages 37–102. Amsterdam/Philadelphia: John Benjamins.
- Miller, Georges A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4):335–391.
- Namer, Fiammetta. 2002. Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Actes de la 9e conférence annuelle sur le traitement automatique des langues naturelles (TALN-2002)*, pages 235–244. ATALA, Nancy.
- Namer, Fiammetta. 2009. *Morphologie, lexique et traitement automatique des langues : L'analyseur DériF*. Paris: Hermès Science-Lavoisier.
- Namer, Fiammetta. 2013a. Adjectival bases of French *-aliser* and *-ariser* verbs: syncretism or under-specification? In N. Hathout, F. Montermini, and J. Tseng, eds., *Selected Proceedings of the 7th Décembrettes: Morphology in Toulouse*,, pages 185–210. München: Lincom Europa.
- Namer, Fiammetta. 2013b. A rule-based morphosemantic analyzer for French for a fine-grained semantic annotation of texts. In C. Mahlow and M. Piottrowski, eds., *SFCM 2013*, CCIS 380, pages 93–115. Heidelberg: Springer.
- Namer, Fiammetta, Pierrette Bouillon, Evelyne Jacquay, and Nilda Ruimy. 2009. Morphology-based enhancement of a French SIMPLE lexicon. In N. Calzolari, A. Rumshisky, P. Bouillon, and K. Kanzaki, eds., *5th International Conference on Generative Approaches to The Lexicon*, pages 153–161. Pisa: ILC-CNR.
- New, Boris, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers* 36(3):516–524.
- Pala, Karel and Dana Hlaváčková. 2007. Derivational relations in Czech WordNet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 75–81. Association for Computational Linguistics.
- Plag, Ingo. 1999. *Morphological productivity. Structural constraints in English derivation*. Berlin/New York: Mouton de Gruyter.
- Plénat, Marc. 2008. Le thème L de l'adjectif et du nom. In J. Durand, B. Habert, and B. Laks, eds., *Actes du Congrès Mondial de Linguistique Française (CMLF-2008)*, pages 1613–1626. ILF, Paris.
- Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Roché, Michel. 2010. Base, thème, radical. *Revue linguistique de Vincennes* 39:95–134.
- Roché, Michel. 2011. Quelle morphologie ? In Roché et al. (2011), pages 15–39.
- Roché, Michel, Gilles Boyé, Nabil Hathout, Stéphanie Lignon, and Marc Plénat. 2011. *Des unités morphologiques au lexique*. Paris: Hermès Science-Lavoisier.

- Romary, Laurent, Susanne Salmon-Alt, and Gil Francopoulo. 2004. Standards going concrete : from LMF to Morphalou. In M. Zock and P. Saint-Dizier, eds., *COLING 2004 Enhancing and using electronic dictionaries*, pages 22–28. Geneva: COLING.
- Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petrucc, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute. Distributed with the FrameNet data.
- Sagot, Benoît, Lionel Clément, Éric Villemont de La Clergerie, and Pierre Boullier. 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of LREC 2006*. Genoa, Italy.
- Sagot, Benoît and Darja Fišer. 2008. Building a free French WordNet from multilingual resources. In *Proceedings of the Ontolex 2008 Workshop*. Marrakech, Morocco.
- Sajous, Franck, Nabil Hathout, and Basilio Calderone. 2013. GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 285–298. Les Sables d’Olonne, France.
- Spencer, Andrew. 1988. Bracketing paradox and the English lexicon. *Language* 64:663–682.
- Stroppa, Nicolas and François Yvon. 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.
- Tanguy, Ludovic and Nabil Hathout. 2002. Webaffix : Un outil d’acquisition morphologique dérivationnelle à partir du Web. In J.-M. Pierrel, ed., *Actes de la 9e conférence annuelle sur le traitement automatique des langues naturelles (TALN-2002)*, pages 245–254. ATALA, Nancy.
- Tribout, Delphine. 2010. *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris 7.
- Tribout, Delphine. 2012. Verbal stem space and verb to noun conversion in French. *Word Structure* 5(1):109–128.
- Tsatsaronis, George and Vicky Panagiotopoulou. 2009. A generalized vector space model for text retrieval based on semantic relatedness. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 70–78. Association for Computational Linguistics.
- Turney, Peter and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 32:141–188.
- Vossen, Piek, ed. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
- Watts, Duncan J. and Steven H. Strogatz. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393:440–442.

- Zeller, Britta D, Jan Snajder, and Sebastian Padó. 2013. DErivBase: Introducing 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.