Morphology-based Enhancement of a French SIMPLE Lexicon

BOUILLON, Pierrette, et al.

Abstract

In this paper, we propose a semi-automatic methodology for acquiring a French SIMPLE lexicon based on the morphological properties of complex words. This method combines the results of the French morphological analyzer DériF with information from general lexical resources and corpora, when available. It is evaluated on a set of neologisms extracted from Le Monde newspaper corpora.

Reference

Morphology-based Enhancement of a French SIMPLE Lexicon

Fiammetta Namer
MSH Lorraine/ATILF
CNRS/Université Nancy2
namer@univ-nancy2.fr

Pierrette Bouillon
ETI / TIM / ISSCO, Geneva
pierrette.bouillon@unige.ch

Evelyne Jacquey
MSH Lorraine / ATILF
CNRS, Nancy
ejacquey@atilf.fr

Nilda Ruimy
ILC-CNRS, Pisa
nilda.ruimy@ilc.cnrs.it

Abstract

In this paper, we propose a semi-automatic methodology for acquiring a French SIMPLE lexicon based on the morphological properties of complex words. This method combines the results of the French morphological analyzer DériF with information from general lexical resources and corpora, when available. It is evaluated on a set of neologisms extracted from Le Monde newspaper corpora.

1 Introduction

There are still no large lexica in Generative Lexicon format (GL, Pustejovsky, 1995), especially for French; we can give three main reasons for this. First, it is difficult to build large scale semantic resources in a systematic way, using well-defined guidelines, general enough to cover a large amount of data. It is also challenging to gather manually all the information necessary for a GL lexicon. Moreover, experience shows that an a priori built lexicon is not very useful for real applications (Bouillon et al., 2000; Bouaud, 1997). The central question is thus: how is it possible to extend existing generative lexica to specialized domains, and keep them updated?

The work described here directly addresses these issues. We propose a semi-automatic methodology for acquiring a French GL lexicon based on the morphological properties of complex words. This method combines the results of the French morphological analyser DériF (Namer, 2002; Namer et al., 2007; Namer, to appear) with information from general lexical resources (TLF for French, Bernard et al., 2002) and corpora, when available. In this way, it tackles the problems mentioned below. First, all words with the same morphological structure receive the same GL representation, which ensures the global coherence of the resource. Second, this methodology makes it possible to extend an a priori lexicon from any given corpus. The DériF analyser starts from any bag of words and, if they are morphologically complex\(^1\), it assigns them some semantic information which is directly exploitable for building a GL resource. This method only applies to complex words, but we know from Cartoni (2008) that they represent a very large proportion of the set of unknown words.

In this paper, we apply our methodology to the semi-automatic extension of a French lexicon based on the SIMPLE model (Lenci et al., 2000) that today constitutes the best example of a generative lexical model. In the following, we present the SIMPLE model, and we describe in detail the proposed methodology for the semi-automatic acquisition of lexical entries. Our main concern is to show how morphology can contribute to the construction of a GL lexicon. We then exemplify and evaluate the methodology for the special case of -eur-suffixed deverbal nouns.

2 SIMPLE

SIMPLE represented a most important European initiative aiming at the design of a standardized model for the creation of rather large-size\(^2\), uniformly structured monolingual semantic lexicons for 12 languages of the European Union, among which some as different as Finnish, Greek or Portuguese.

The SIMPLE semantic model, consensually adopted by all European partners of the SIMPLE project, imposed itself as a de facto standard and later strongly inspired the Lexical Markup Framework, which is now the ISO standard\(^3\) for NLP lexicons. That model, built with a view to multilinguality, aimed at achieving a high level of harmonization among the semantic lexicons for the different languages.

The level of semantic representation was added on top of the 12 morphological and syntactic lexicons previously elaborated in the framework of the PAROLE European project; the candidates to semantic description were selected among the words encoded at those two levels and according

---

1 For space reasons, morphologically complex words will henceforth be referred to as complex words.
2 10,000 word meanings.
3 ISO-24613:2008
to their frequency in the PAROLE corpus. Moreover, in order to guarantee an overlapping of senses across languages as well as a uniformity of coverage throughout the different semantic types, a common set of EuroWordNet base concepts was encoded in all languages.

The SIMPLE model, which builds on the results of outstanding European projects such as EAGLES, GENELEX, ACQUILEX and EUROWORDNET, allowed to create large repositories of generic and explicit lexical information with variable degrees of granularity. Besides the theoretical and representational model, these lexicons share a common building methodology, a data management tool, a DTD and the XML output format. Thanks to their high degree of genericity, modularity and coherent structuring, the SIMPLE lexical resources lend themselves to extension, reusability, customization and tuning to meet the requirements of different NLP applications.

The theoretical framework underlying the SIMPLE model is the Generative Lexicon (Pustejovsky, 1995, 2001). In a generative lexicon, a lexical unit is modelled through four different levels of representation⁴ that account for the componential aspect of meaning, define the type of event denoted, describe its semantic context and set its hierarchical position with respect to other lexicon units.

SIMPLE semantic lexicons are structured in terms of the SIMPLE core ontology that moves from a GL basic assumption, viz. word senses are multidimensional entities with different degrees of internal complexity: some may be exhaustively described through a taxonomical relation, whereas, for the characterization of others, orthogonal dimensions of meaning come into play. Accordingly, the 157 language-, domain- and application-independent semantic types that make up the SIMPLE ontology are organized on the basis of orthogonal principles (Pustejovsky and Boguraev, 1993), and the multidimensionality of meaning is captured by means of the four roles of the Qualia Structure. Besides, in the SIMPLE ontology, semantic types are not taken as mere labels but rather as bundles of structured semantic information (the type-defining and the optional information characterizing the semantic type INSTRUMENT are shown in Table 6). Assigning a semantic type to a lexical unit is therefore tantamount to endowing it with the set of properties characterizing this type. The defining properties of each semantic type are collected in a template (cf. Table 6), i.e. an underspecified, schematic structure that sets the well-formedness requirements for semantic units candidate to membership. The template-based lexicon building methodology, which ensured language-internal and cross-language uniformity and coherence among the 12 SIMPLE lexical resources, proves particularly helpful in the present work for deriving lexical entries from the information provided by DériF (see section 4).

The Qualia Structure provides a formal language to model the information regarding the different semantic components that contribute to defining the internal structure of a lexical unit — whatever its syntactic category — as well as its relations with other units of the lexicon. In designing the SIMPLE model, this structure was revisited with a view to enhance its expressive power; this gave rise to the Extended Qualia Structure whereby each of the four original roles subsumes a set of subtypes expressed in terms of relations between semantic units.

Each predicative lexical unit of the lexicon is related, through a defined type of link⁵, to a lexical predicate which is described in terms of semantic role and selectional restrictions of its arguments. Semantic and syntactic information are then correlated through the mapping of the argument structure onto the syntactic frame.

A semantic unit, which represents a unique meaning of a lexeme, is therefore endowed with the following range of information formally expressed as weighted semantic features or relations (cf. Table 7, for a partially filled semantic entry):

- Semantic type
- Domain of use
- Definition and/or example
- Event type: state, process or transition
- Logical polysemy
- Synonymy relation
- Derivation relation
- Extended Qualia Structure relations
- Argument structure, typing of args.
- Link semantic/syntactic representation

SIMPLE lexical resources were meant as core lexicons to be further extended, and actually some of them were then enlarged in the context

---

⁴ Argument Structure, Event Structure, Qualia Structure and Lexical Inheritance Structure.

⁵ master, agent_nominalization, patient_nominalization, etc.
of national projects. To our knowledge, the French SIMPLE lexicon, which was derived from LexiQuest's French lexicon, did not undergo any extension or updating in the SIMPLE format. In this context, our current initiative seems therefore all the more sound and timely.

3 Extending the semantic lexicon

The lexicon extension method presented here takes advantage of the informative potential that derivational morphology provides through constraints rules exert on both bases and complex word they link together. Precisely, the related experiment has been carried out on 338 new coined nouns extracted from Le Monde newspaper corpus. These nouns, which all end with -eur, are therefore seemingly suffixed words. The choice of adding neologisms to the current SIMPLE lexicon is a reasonable guaranty against the non-compositionality of complex meaning: words that have been recorded in dictionaries a long time ago often bear both an opaque meaning and frozen characteristics that morphology is no longer able to reveal. On the contrary, new coined words have a predictable definition and regular semantico-syntactic features. The approach described in this section is based on results produced by the morpho-semantic parser DériF (Namer, 2002; to appear).

A word formation rule (WFR) is connected to prototypical semantic, syntactic and phonological constraints that apply on both base and complex words linked by the WFR. Knowing these constraints is an asset for NLP since the most productive and regular ones are reused in order to serve as input for the automatic acquisition of lexical features, and thus to enhance the lexicon content to be translated into the SIMPLE format. In what follows, we will see what kind of semantic information DériF is capable of acquire (3.1); then the automatic assignment of morpho-semantic features is illustrated through the analysis of -eur suffixed nouns (3.2). Afterwards, the application of internal specification and unification methods on these features is described (3.3).

Finally, section 3.4 explains the way dictionary content is used to disambiguate these features.

3.1 Acquisition via morphological parsing

DériF methodology is based on the application of an analysis chain. Each step is a rule applying on a part-of-speech tagged lemma. In case of derivation, the rule links the lemma to its morphological base (i.e. another lemma). While doing so, the rule also provides the analysed lemma with its linguistic meaning defined with respect to its base. This meaning is expressed by means of a gloss in natural language. In turn, the calculated base is considered as a possible complex word to be analysed and serves as a new DériF's input, and so on, until the obtained base is a simplex word. Besides this functional aspect, DériF has other special features. For instance, when a word is morphologically ambiguous (e.g. implantable, establishable / unplantable), DériF provides it with all possible analyses (one produces plantable, the other one implantable). Moreover, a default analysis (provided when each other specific operation fails) is systematically proposed by each rule, and this accounts for the morphological regularity of complex neologisms. Finally, in addition to the gloss defining the analysed word with respect to its base (illustrated in Table 1, line 1), DériF assigns to both words other lexical information (Namer et al., 2007), as detailed below. This information mirrors the constraints the WFR involves when linking these words. The way these features are automatically assigned is illustrated through the example of -eur deverbal nouns formation.

3.2 Semantic tagging: -eur deverbal nouns

French -eur suffixed deverbal nouns are usually described as agentive nouns, (Scalise, 1984; Corbin, 2001). More precisely, the agentive (chanteurs ‘singer’) or instrumental interpretation (interrupteurs ‘switch’) is conditioned by the mandatory presence of an agentive (dynamic) base verb (Busa, 1997; Fradin, 2003; Kerleroux, 2004). Very rare exceptions to this principle (e.g. naissseur ‘born-er’, trebucheurs ‘stumble-er’) correspond most of the time to a causative reading (naissseur ‘He who makes someone be born’) (Fradin et al., 2003). As for -eur nouns, they refer to concrete entities, the animate (for agents) or non-animate nature (for instruments) of which is not morphologically predictable. The knowledge

---

6 We assume a word to be a neologism when it is not included in the French reference dictionary TLF (Trésor de la Langue Française), which contains more than 90,000 nouns, verbs and adjectives belonging to the general language. Often, neologisms found in newspapers belong to specialized languages, such as stock-exchange (arbitrageur ‘arbitration-er’) or media (audimateur ‘audience-rating-er’) terminologies.

7 In this paper, lexemes are noted in italics; SIMPLE attributes names in small capitals.
of these theoretical assumptions allows DériF to assign three features sets to the noun/verb pairs linked by the -eur WFR (cf. Table 1)\(^8\):

1. the noun meaning, defined according to the base verb V value and expressed as: ‘agent - instrument of V’ (line 1);
2. constraints about the semantic restriction on -eur nouns (line 3): they refer either to human beings (hum=yes) or to artefacts (natural=no, anim=no, concrete=yes); constraints about the base verb aspectual value (aspect = dynamic) and about its agentivity (sub-cat=<NPatient,...>) (line 2)\(^3\);
3. expected predicate-argument relations, between the -eur noun and its base verb: the former is indexed with @2, and this index matches the agent subject NP subcategorized by the base verb, indexed with @1 (line 4).

|   | caramélisateur/NOUN==>
|---|---------------------------|
| 4 | VERBE/eur/suf/NOM+caraméliser/VERBE
|   | NOM/iser/suf/VERBE+caramel/NOM; "(Usual agent - Occasional author - Instrument) ofcaraméliser" \(\theta\)

|   | caraméliser/VERB: @1 [ aspect = accomplissement, subcat = <NPagent, ...> ]
|   | caramélisateur/VERB: @2 [ aspect = dynamic, subcat = <NPagent, ...> ]
| 3 | caramélisateur /NOUN: @2 [ concrete = yes, hum = yes, count = yes, subcat = @2 [ concrete = yes, anim = no, natural = no, count = yes ] \(\delta\), rel = NPagent( @2, sub_cat( @1 )) ]
| 4 | rel = NPagent( @2, sub_cat( @1 )) \(\Theta\) |

Table 1. Tagging caramélisateur and caraméliser.

### 3.3 Internal Disambiguation

Annotations as those illustrated in lines 2-4 in Table 1 above are assigned by DériF during each step of a complex noun Neur analysis process. Consequently, when the Neur base verb is itself morphologically complex, this verb is likely to receive two features sets. One reflects its being the Neur base (cf. line 2); the other records the features the verb gets as output of another WFR. This is the case for caraméliser, since this verb is derived by the -iser WFR from the noun caramel (caraméliser: ‘Turn <-{Part-of}obj> into caramel’). Now, this WFR produces a twofold substitution process on DériF annotation results. Precisely, the annotation reproduced in Table 3 substitutes for, and is more precise than, both features sets that characterize caraméliser on line 2, Table 1, and in Table 2.

|   | caraméliser/VERB: [ aspect = accomplissement, subcat = <NPagent, NPatient > ]

Table 2. Tagging caraméliser as derived from caramel

In brief, the morphological parsing of caramélisateur leads DériF to provide caraméliser with two different set of features. This multiple information has to be unified: this task is performed on DériF output by a cross-validation filter which specifies, complete or sometimes redefines the competing contents. For our example, the unification task allows to predict that the verb (1) denotes an accomplishment, (2) is agentive (3) (consequently) is strictly transitive. This disambiguation leads to a twofold substitution process on DériF annotation results. Precisely, the annotation reproduced in Table 3 substitutes for, and is more precise than, both features sets that characterize caraméliser on line 2, Table 1, and in Table 2.

|   | caraméliser/VERB: [ aspect = accomplissement, subcat = <NPagent, NPatient > ]

Table 3. Cross-validation: caraméliser final features

Moreover, indices @1 and @2 are recreated in order to preserve the relationships the verb keeps with its base noun on the one hand, and with the -eur noun on the other hand.

Though the cross-validation filter has been presented with -eur deverbal nouns / -iser denominal verbs interaction, it is activated with many more morphological verb structures. For instance, when an -eur deverbal noun is analysed, the disambiguation filter is set off when the following nominal, deverbal or deadjectival verb formation rule is identified (Table 4): -iser and -ifier suffixation, en-, a-, dé-, é- prefixation, conversion. Clearly, the verb final features value varies according to the specific constraints each rule involves.

- **-iser**
  - mobile\(\alpha\) > mobiliser > mobilisateur
  - ampli\(\alpha\) > amplifier > amplificateur
  - momi\(\alpha\) > momifier > momificateur
  - en-
  - joli\(\alpha\) > enjoliver > enjoliveur
  - pail\(\alpha\) > empailleur > empaillleur
  - a-
  - grand\(\alpha\) > agrandir > agrandisseur
de-
  - boucher\(\alpha\) > déboucher > déboucheur
  - crasse\(\alpha\) > décrasser > décrasseur
  - é-
  - grap\(\alpha\) > égrappager > égrappeur
  - conv\(\alpha\) > vider > videur
  - conv\(\alpha\) > balayer > balayeur

\(^{8}\) Black numbers in Table 1 and Table 3 are taken up in section 4 and in Table 7.

\(^{3}\) The three dots (...) indicate the unpredictability of the -eur noun base verb transitivity, excepted for the agentive nature of the subject.
Table 4. Cases of *Neur* base-verb double tagging

### 3.4 Contribution from dictionaries

#### 3.4.1 Motivations

As said before, morphological constraints take into account predictable information for complex words at the semantic, syntactic and phonological levels. However, they cannot, and are not expected to, handle lexical information that is related to the uses of these complex words. Given the choice to study neologisms, that is words with a compositional meaning, one could decide that morphology is sufficient. However, even for neologisms, some lexical information may lack or remain underspecified. In order to define usage rules, we first established and assessed an extraction methodology on existing complex words and then we applied it to neologisms.

Dictionaries may address such an issue because they are especially intended to record usage information of complex and simple words in a language. That is why we decided to extract useful information from one of the best and freely exploitable dictionaries for French, the TLFi (Trésor de la langue française informatisé, Bernard and al., 2002).

#### 3.4.2 Semi-automatic extraction and results

The methodology and results presented here are illustrated with *Neur* nouns. Recall that their animate (for agents) or non-animate nature (for instruments) is not morphologically predictable (see section 3.2).

Our methodology aims, among others, to solve this semantic ambiguity: it takes advantage from the fact that lexicographic definitions in French dictionaries are built following the denotations of nouns with respect to the human / artefact distinction. In TLFi for example, the definition of *chanteur*, ‘singer’ begins with *Celui qui chante* ‘the one who sings’ and the more general definition of *amortisseur*, ‘shock absorber’ begins with *Dispositif qui atténue la violence de quelque chose* ‘device which makes the violence of something decrease’. Agentive definitions may also be revealed by some typical words denoting human agents like *Ouvrier*, ‘worker’, *Employé*, ‘employee’, *Homme*, ‘man’, etc. In brief, a *Neur* definition of the form ‘*<Pro-HUM>* / *<Hum-Agent>* who *VERB*’, which includes an ‘agent marker’, will lead to an agentive interpretation for *Neur*, whereas ‘tool/device/ artefact... which *VERB*’ serves to assign the *Neur* noun an instrumental reading and starts with any ‘instrumental marker’.

Given such observations, we built a specific modularized and incremental parser which takes a decision among a threefold choice: (1) the given *Neur* noun refers to an agent, (2) to an instrument or (3) either to an agent or an instrument (AMB in Table 5) when the parser encounters both agentive and instrumental definitions.

In a first step, we assessed the parser results with the 2,258 *Neur* nouns recorded in the TLFi and out of which Dérif selects 1,683 complex nouns, that is ~75% (the remaining 25% are simple nouns). In this subset, recall and precision (Manning and Schütze, 2002) have been manually calculated by human examination of the parser result.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Agent</th>
<th>Instrument</th>
<th>Amb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,259 (75%)</td>
<td>193 (11%)</td>
<td>231 (14%)</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>1</td>
<td>1</td>
<td>0.70</td>
</tr>
<tr>
<td>Precision</td>
<td>0.93</td>
<td>0.94</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5: Interpretations for *Neur* from TLFi

Among the 1,683 *Neur* matching the decision criteria, 75% are tagged by the parser as agents, 11% as instrument, and 14% as both. Human validation shows that 83/1,259 nouns are wrongly tagged as agent and that 12/193 of them are wrongly tagged as instruments, but none is wrongly tagged as both. Moreover, wrong parser decisions are replaced by ambiguous tagging - that is both instrument and agent. This explains the low recall score (0.70) for AMB tagging. Finally, no complex deverbal -eur noun is ignored by the parser criteria, which explains the perfect recall scores for agent and instrument.

Given such results on existing complex deverbal -eur nouns, a similar methodology was adopted for neologisms.

#### 3.4.3 Application to neologisms

For neologisms, the problematic is a bit different and the parser will have to be adapted. Two different cases should be distinguished. Either they can be found in other dictionaries on Internet or it is possible to exploit the context in which they are used in the corpus. In the former case, the methodology described above is directly applicable since definitions follow the rules used for the TLFi. This the case with “ambianceur” or “affabulateur”. In the later case, two strategies can be used for disambiguation: (1) the *Neur* noun occurs into a coordinate structure or an enumeration whereby an already disambiguated -eur noun occurs too; (2) the *Neur* noun occurs in a disambiguating context, that is a con-
text which contains some unambiguous lexical elements (to be employed or to exercise a given profession, etc. for human agents; to have a price, to be bought, etc. for artifacts). For example, “Il était ce qu'on appelle un aviseur, un informateur en quelque sorte, mais bien particulier” (Google search, June 2009) contains both an attributive structure with the ambiguous neologism aviseur and an enumeration with informateur. Since informateur, a noun existing in the TLFi was automatically tagged as an agent, we may conclude that aviseur is an agent too. This methodology will be evaluated in Section 5 on neologisms extracted from Le Monde corpus.

4 From DériF to SIMPLE: the case of -eur ending complex nouns

At the end of the analysis and disambiguation process, the DériF morphological analyzer outputs a rich set of information, among which some semantic features. In this section, we show how this information is used to derive semi-automatically French lexical entries for -eur ending complex nouns, in compliance with the SIMPLE model. According to the SIMPLE ontology, deverbal nouns ending in -eur are classified under two main type hierarchies: either HUMAN or ARTIFACT. Under the first hypothesis, they may belong to three different semantic types: AGENT_OF_TEMPORARY_ACTIVITY, characterized by an agentive relation, i.e. either ‘agentive’ (killer, to kill) or ‘agentive_prog’ (walker, to walk); AGENT_OF_PERSISTENT_ACTIVITY, defined by the telic relations ‘is the ability of’ (skier, to ski) or ‘is the habit of’ (smoker, to smoke); PROFESSION whereby the telic role is expressed through the relation ‘is the activity of’ (tiler, to tile). Under the second hypothesis, they belong to the type INSTRUMENT, characterized by both an agentive and a telic meaning dimensions, as shown in Table 6. In Tables 1 and 3, the DériF features relevant to the encoding of the SIMPLE lexical entry for caramélisateur are, in particular, the derivational base of the processed complex word 1; the semantic relationship holding between the noun and its verbal base 2; the verb argument structure 3; some semantic properties of the noun 4 and the argument the noun lexicalizes 5. Using these features, the semi-automatic derivation of SIMPLE entries consists in a two-stage process. As mentioned in section 2, the SIMPLE model advocates a template-based encoding strategy. The first stage consists therefore in using the information given in 2 for selecting, within the SIMPLE ontology, the appropriate semantic type to be assigned to the lexical unit. ‘Usual agent’ suggests a membership in two possible semantic types, viz. AGENT_OF_PERSISTENT_ACTIVITY and PROFESSION, while ‘Exceptional/Occasional author’ points to AGENT_OF_TEMPORARY_ACTIVITY and ‘Instrument’ to INSTRUMENT.

Following the semantic type selection, the encoding process goes on with the instantiation of the corresponding template. After information 2 is correctly disambiguated (see section 3.4.2), the template INSTRUMENT (Table 6) is instantiated and its underspecified information is filled with the features provided by the DériF analyzer, as shown in Table 7. Some DériF features are directly usable, others need to be transformed into the SIMPLE format. At the end of the migration process, the features synergy with the information given by the template in order to produce a well-formed SIMPLE entry to be finalized then with the selectional restrictions of the predicate’s arguments and some optional information. In the following, we evaluate this methodology for extending the French SIMPLE lexicon with -eur complex nouns.

![Table 6. Template INSTRUMENT](image)

<table>
<thead>
<tr>
<th>Usem:</th>
<th>caramélisateur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template_Type:</td>
<td>[INSTRUMENT]</td>
</tr>
<tr>
<td>Unification_path:</td>
<td>[Concrete_entity</td>
</tr>
<tr>
<td>Domain:</td>
<td>General</td>
</tr>
<tr>
<td>Gloss:</td>
<td>//free//</td>
</tr>
<tr>
<td>Predicative:</td>
<td>Pred_caraméliser: Arg1agent,Arg2patient</td>
</tr>
</tbody>
</table>

![Table 6. Template INSTRUMENT](image)
5 Evaluation

Among the 338 initial -eur ending novel nouns found in Le Monde, 33 are correctly identified as simplex words with respect to the -eur WFR (for instance, métamoteur ‘meta-engine’). 96 others are not accounted for in this evaluation since they are not neologic. They are both suffixed with -eur and either prefixed or compound (in other words, -eur WFR is not the last applied rule: e.g. euroconsommateur ‘euroconsumer’ is a clipped compound meaning ‘European consumer’).

In order to assess the migration of DériF analysis output into information relevant for populating SIMPLE-compliant entries, we considered only the 210 nouns analysed as complex words by DériF and disambiguated into AGENT, INST(rument) and AMB(iguous), following rules explained in 3.4.3. Among these 210 nouns, 18% (37) were found in other dictionaries on Internet (for example ambiencieur or affabulator) and 73 % (173) appeared in corpora. In total, 193 were disambiguated following rules described in 3.4.3 : 156 were tagged as AGENT, 29 as INST and 8 as AMB while 9% (17) remained unsolved by the proposed rules.

Once the words were disambiguated, DériF information filled the selected appropriate template, as illustrated in section 4. INST tagged words were assigned a unique SIMPLE template, viz. INSTRUMENT, whereas AGENT tagged words fell into three possible description frames, namely the templates PROFESSION, AGENT_OF_PERSISTENT_ACTIVITY and AGENT_OF_TEMPORARY_ACTIVITY. AMB-tagged words and words without tags (unsolved), in turn, were assigned the four above-mentioned templates.

The number of neologisms with at least one correct SIMPLE template assignment — which was then calculated — is 92% (193/210). Only 29/210 words were assigned one single template. Others will be associated to a minimum of three templates, among which at least one is correct. Even for those words that cannot be automatically disambiguated yet, the automatic procedure is still convenient in terms of time economy with respect to the full manual encoding of a lexical entry.

Unsolved cases are due to different factors: i) prototypicity of the DériF analysis (-eur agentive nouns refer primarily to human beings, however, e.g. nicheur, although being an agent, only denotes animals, for pragmatic reasons (typical agents for the ‘nesting’ activity are animals). Other nouns that are neither artefact or human include attracteur (a mathematical concept in the corpus); ii) erroneous word segmentation and misspellings (ex: aileur or ailleurs vs. ailleurs); iii) wrong PoS tagging, e.g.: diabolisateur is an adjective in the corpus but is tagged as a noun. These results are deemed positive all the more since cases of i) are very rare (6/210). For the most frequent error type, i.e. iii) (11/210), a solution could consist in detecting more accurately Adj and N uses and restricting the disambiguation to Neur uses only.

6 Conclusion

In this paper, we have shown how three types of knowledge could be combined in order to produce a lexical resource exploitable in NLP. In fact, the principles of lexical morphology, lexicographic definitions freely available and the lexical semantics model SIMPLE were all brought into play with a view to extending a French semantic lexicon built in an operational format. The assessment of the results of this experiment, carried out on a sample of 338 nouns extracted from Le Monde corpus but missing from dictionaries, shows that it is very conclusive.

The methodology is straightforwardly generalizable to the whole set of 3,301 -eur suffixed complex nouns encoded in the TLF.

<table>
<thead>
<tr>
<th>Representation:</th>
<th>caramélisateur : agent_nominalisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection Restr.:</td>
<td></td>
</tr>
<tr>
<td>Derivation:</td>
<td>deverbaleNounVerb</td>
</tr>
<tr>
<td></td>
<td>&lt;caramélisateur, caraméliser&gt;</td>
</tr>
<tr>
<td>Formal:</td>
<td>isa (caramélisateur, instrument)</td>
</tr>
<tr>
<td>Agentive:</td>
<td>created by (caramélisateur, &lt;fabriquer [CREATION]&gt;)</td>
</tr>
<tr>
<td>Constitutive:</td>
<td>made_of (1, &lt;Usem&gt;) //optional/</td>
</tr>
<tr>
<td></td>
<td>has_as_part (1, &lt;Usem&gt;) //optional/</td>
</tr>
<tr>
<td>Telic:</td>
<td>used_for (caramélisateur, &lt;caraméliser [EVENT]&gt;)</td>
</tr>
<tr>
<td>Synonymy:</td>
<td>Synonym (1, &lt;Usem: [Instrument]&gt;) //optional/</td>
</tr>
<tr>
<td>Reg. Polysemy:</td>
<td>&lt;Nil&gt;</td>
</tr>
</tbody>
</table>

Table 7. DériF information-based automatic encoding
In the short term, other morphological types of complex nouns could undergo the same process to enrich the French SIMPLE lexicon with new entries and additional features.

In the medium term, the information predicted by DériF on complex adjectives and verbs (see section 3.3, Tables 2 and 3), and which cannot be expressed in the SIMPLE formalism for the time being will be studied in order to integrate this relevant part of the lexicon into the whole set of SIMPLE resources. It is in this sense that morphology contributes to validate and refine the SIMPLE entries.

References


