HAL open science

Morphology-based Enhancement of a French SIMPLE Lexicon

Fiammetta Namer, Pierrette Bouillon, Evelyne Jacquey, Nilda Ruimy

To cite this version:

Fiammetta Namer, Pierrette Bouillon, Evelyne Jacquey, Nilda Ruimy. Morphology-based Enhancement of a French SIMPLE Lexicon. 5th International Conference on Generative Approaches to the Lexicon, Sep 2009, Pise, Italy. pp.1-8. hal-00413324

HAL Id: hal-00413324

https://hal.science/hal-00413324v1

Submitted on 3 Sep 2009

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

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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-CNR, Pisa
nilda.ruimy@ilc.cnr.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 (XXX 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 (XXX, 2002; XXX et al., 2007; XXX, 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, ensuring 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 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), which 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 the first 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 (Lenci et al., 2000), 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. This 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

[1] For space reasons, morphologically complex words will henceforth be referred to as complex words.
[2] 10,000 word meanings.
[3] ISO-24613:2008

the PAROLE European project; the candidates to semantic description were selected among the words encoded at these two levels and according 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 in order 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 representation[4] 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 SIMPLE ontology, semantic types are not taken as mere labels but rather as bundles of structured semantic information. 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*, 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 link[5], 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:

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

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

[5] master, agent_nominalization, patient_nominalization, etc.

SIMPLE lexical resources were meant as core lexicons to be further extended, and actually some of them were then enlarged in the context of national projects. To our knowledge, after the end of the SIMPLE project, 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 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[6] extracted from *Le Monde* newspaper corpora. 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, which 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 (XXX, 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 categorised lemma. In case of derivation, the rule links the analysed lemma to its morphological base (i.e. another categorised lemma). While doing so, the rule also provides the analysed lemma with its linguistic meaning defined wrt its base. This pseudo-definition 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*$_A$[7] 'establishable / unplantable'), DériF provides it with all possible analyses (one produces *plantable*$_A$ 'plantable', the other one *implanter*$_V$ 'establish'). Moreover, a default analysis 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 wrt its base (illustrated in Table 1, line 1), DériF assigns to both words other lexical information (XXX et al., 2007). 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 (*chanteur*$_N$ 'singer') or instrumental interpretation (*interrupteur*$_N$ '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. *naisseur*$_N$ 'born-er', *trébucheur*$_N$ 'stumble-er') correspond most of the time to a causative reading (*naisseur*$_N$ = '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.

[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):

(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 being (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=<NPagent,...>) (line 2)[8];

(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).

1	caramélisateur/NOUN==> 4,VERBE/eur/suf/NOM+caraméliser/VERBE❶ NOM/iser/suf/VERBE+caramel/NOM": "(Usual agent - Occasional author - Instrument) ofcaraméliser" ❷
2	caraméliser/VERB: @1 [aspect = dynamic, subcat = <NPagent, ...>]
3	caramélisateur /NOUN: @2 [concrete = yes, hum = yes, count = yes] ‖ @2 [concrete = yes, anim = no, natural = no, count = yes] ❹;
4	rel = NPagent(@2, sub_cat(@1)) ❺

Table 1. Tagging *caramélisateur* and *caraméliser*

3.3 Internal Disambiguation

As we said, 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 accomplishment, change-of-state or change-of-location verbs, therefore verbs that always select a patient argument. Sometimes, *-iser* verbs realize strictly unaccusative predicates (*fraterniser* 'fraternize'), sometimes they also select an agent/cause (*atomiser* 'atomize'), and sometimes they accept a causative-unaccusative alternation (*caraméliser*). Owing to this triple potential argument structure, DériF assigns to *-iser* ending denominal verbs a subcategorisation list with an optional agent; expressed by a parentheses notation (subcat = < (NPagent), NPpatient>).

caraméliser/VERB: @1 [aspect = accomplissement, subcat = <(NPagent), NPpatient >]

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, NPpatient >] ❸

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 denominal, 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	$mobile_A$ > mobiliser > mobilisateur
-ifier	$ample_A$ > amplifier > amplificateur
	$momie_N$ > momifier > momificateur
en-	$joli_A$>enjoliver>enjoliveur
	$paille_N$ > empailler>empailleur
a-	$grand_A$>agrandir>agrandisseur
dé-	$boucher_V$> déboucher > déboucheur
	$crasse_N$ > décrasser > décrasseur
é-	$grappe_N$ > égrapper>égrappeur
$conv_{A>V}$	$vide_A$>vider>videur
$conv_{N>V}$	$balai_N$>balayer>balayeur

Table 4. Cases of ***Neur*** base-verb double tagging

[8] The three dots (...) indicate the unpredictability of the *-eur* noun base verb transitivity, excepted for the agentive nature of the subject.

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 wrt human / artefact distinction. In TLFi for example, the definition of *chanteur*$_N$ 'singer' begins with *Celui qui chante* 'the one who sings' and the more general definition of *amortisseur*$_N$ '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 have built a modularized and incremental parser which takes a decision among a threefold choice: (1) the given ***Neur*** noun refers to an agent, (2) it refers to an instrument or (3) it may either refer to an agent or an instrument (AMB in Table 5) when the parser encounters both agentive and instrumental definitions.

In a first step, we have assessed the parser results with the 2.258 ***Neur*** nouns recorded in the TLFi.

	Neur (2,258)	AGENT	INSTR	AMB
Recall	1,719 (76%)	1,235 (72%)	197 (11%)	287 (17%)
Precision	1,719 (100%)	~1,100 (89%)	~170 (85%)	~220 (76%)

Table 5: Interpretations for ***Neur*** from TLFi

As shown in this table, first line, 76% of the 2,258 ***Neur*** nouns satisfy at least one of the criteria that are used in the parser, but without mistakes (precision 100%). The high precision score is due to the pre-selection performed by DériF. Among the 1.719 ***Neur*** matching decision criteria, 72% are tagged as agents, 11% as instrument, and 17% as both. Precision is, in each case, relatively high.

In the near future, the precision score will be easily increased by means of the incremental conception of the parser. In TLFi, not all definitions are regular, even though it is the case for most of them. Precisely, the so-called 'agent markers' or 'instrument markers' may not be present in definitions. A human validation step will allow to list new disambiguating markers. Moreover, the criteria list will be extended in order to take into account another distinction, the one between usual agents and occasional ones.

3.4.3 Application to neologisms

Our experiment concerns 210 nouns out of the initial set of 338 new coined *-eur* ending nouns, as these 210 nouns are analysed as complex words by DériF. In this subset, 18% of nouns (37) can be found in other dictionaries on the Internet. For this first subset, one observes that all given definitions match the ones given by TLFi. So, in this case, the methodology is directly applicable.

The second subset (82%, 173) can be divided into two cases: (1) the ***Neur*** noun occurs into a coordinate structure or an enumeration one and in this structure, an already disambiguated *-eur* noun occurs too; (2) the ***Neur*** noun occurs in a disambiguating context, that is a context 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 the first subset (***Neur*** nouns found in Web dictionaries), ambiguity disappears, as expected (cf. section 3.1): ***Neur*** nouns are mainly interpreted as agents (70%, 26), then as instruments (30%, 8). For the second set too, human interpretation is preferred (75%, 131). The first difference concerns unsolved cases (9%, 16) and ambiguous cases frequency (3%, 5), which increase against that of instrumental reading (12%, 21). These cases are mainly caused by tagging errors (adjectives vs. nouns, proper nouns, compounds) which could be solved with additional rules. Very few interesting exceptions are three nouns which do not refer to a human-agent or an artefact (*attracteur*, an 'attractor', mathematical concept; *nicheur*, a kind of bird which nests; *dérégulateur*, a kind of substance which 'deregulates' something).

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

At the end of the analysis 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 sky*) 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 lexical entry for *caramélisateur* in the framework of the SIMPLE model are, in particular, the derivational base of the processed complex word ❶; the semantic relationship holding between the noun and its verbal base ❷; the verb argument structure ❸; some semantic properties of the noun ❹ and the argument the noun lexicalizes ❺. 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 ❷ 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 ❷ 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.

Usem:	1
Template_Type:	**[INSTRUMENT]**
Unification_path:	**[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
Domain:	*General*
Gloss:	*//free//*
Predicative Representation:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Derivational relation>
Formal:	*isa* (1, <instrument> or <hyperonym>)
Agentive:	*created_by* (1, <manufacture: [Creation]>)
Constitutive:	*made_of* (1, <Usem>) //optional// *has_as_part* (1, <Usem>) //optional//
Telic:	*used_for* (1, <Usem: [Event] >)
Synonymy:	*Synonym* (1, <Usem: [Instrument]>) //optional//
Reg. Polysemy:	<Nil>

Table 6. Template INSTRUMENT

Usem:	**caramélisateur**
Template_Type:	**[INSTRUMENT] ❷ ❹**
Unification_path:	**[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
Domain:	*General*
Gloss:	*//free//*
Predicative Representation:	**Pred_caraméliser: Arg1agent,Arg2patient ❸** **caramélisateur : agent_nominalisation ❺**
Selectional Restr.:	==
Derivation:	***deverbalNounVerb*** **<caramélisateur, caraméliser> ❶**
Formal:	***isa* (caramélisateur, instrument)**
Agentive:	***created_by* (caramélisateur, <fabriquer [CREATION]>)**
Constitutive:	*made_of* (1, <Usem>) //optional// *has_as_part* (1, <Usem>) //optional//
Telic:	***used_for* (caramélisateur, <caraméliser [EVENT]>) ❷**
Synonymy:	*Synonym* (1, <Usem: [Instrument]>) //optional//
Reg. Polysemy:	<Nil>

Table 7. DériF information-based automatic encoding

5 Evaluation

Among the 338 initial *-eur* ending novel nouns found in *Le Monde*, 33 are correctly identified as simplex words wrt 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 209 nouns analysed as complex words by DériF and disambiguated into AGENT, INST(rument) and AMB(iguous) (as illustrated in 3.4.3). 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, in turn, were assigned the four above-mentioned templates.

The recall rate ─ i.e. the number of neologisms with at least one correct SIMPLE template assignment ─ which was then calculated is 82% (172/209). Failures 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; ii) erroneous word segmentation and misspellings; iii) wrong PoS tagging (word tagged as a noun although it is used as an adjective in the text).

These results are deemed positive all the more since cases of i) are very rare (6/209). For the most frequent error type, i.e. iii) (20/209), 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.

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 in the whole set of SIMPLE resources. It is in this sense that morphology contributes to validate and refine the SIMPLE entries.

References

Bernard P., Lecomte J., Dendien J., Pierrel J.-M. 2002. Un ensemble de ressources informatisées et intégrées pour l'étude du fançais : FRANTEXT, TLFi, Dictionnaire de l'Académie et logiciel Stella, présentation et apprentissage de leur exploitation, TALN 2002.

Bouillon P., Fabre C., Sébillot P., Jacqmin L. 2000. Apprentissage de ressources lexicales pour l'extension de requêtes, TAL 41:2 2000.

Busa, F. 1997. The Semantics of Agentive Nominals in the Generative Lexicon. *Predicative Forms in Natural Language*, ed. P. Saint-Dizier. Amsterdam: Kluwer.

Busa, F. et al. 2001. Generative Lexicon and the SIMPLE Model : Developing Semantic Resources for NLP. Bouillon P. & F. Busa (Eds.) *The Language of Word Meaning*. Cambridge, MA : Cambridge University Press, 333-349.

Calzolari, N. et al. 2003. SIMPLE: Plurilingual Semantic Lexicons for Natural Language Processing. Zampolli A., N. Calzolari & L. Cignoni (Eds.) *Computational Linguistics in Pisa. Linguistica Computazionale*, Special Issue, XVIII-XIX. Pisa-Roma : IEPI. Tomo I, 323-352.

Corbin, D. 2001. Préfixes et suffixes: du sens aux catégories. *Faits de Langue*, 41-69. Paris: Ophrys.

Bouaud J., Habert B., Nazarenko A., Zweigenbaum P. 1997. Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation avec deux modélisations conceptuelle, Acte du Colloque ingénierie des Connaissances, Roscoff, 1997.

Cartoni Br. 2006. Constance et variabilité de l'incomplétude lexicale, Actes de Récital, 661-669.

Fradin, B. 2003. *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.

Fradin, B., and Kerleroux, F. 2003. Quelles bases pour les procédés de la morphologie constructionnelle ? *Silexicales 3 : les unités morphologiques*, eds. B. Fradin et al., 76-84. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Kerleroux, F. 2004. Sur quels objets portent les opérations morphologiques de construction ? *Lexique*:85-124.

Lenci, A., et al. (2000), SIMPLE Linguistic Specifications, Deliverable D2.1, ILC-CNR, Pisa.

Lenci, A. et al. 2000. SIMPLE: A General Framework for the development of Multilingual Lexicons. *International Journal of Lexicography,* special issue, Dictionaries, Thesauri and Lexical-Semantic Relations 13(4), 2000. 249-263.

Namer, F. 2002. Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. *TALN-2002*, Nancy, France:235-244.

Namer, F., Bouillon, P., and Jacquey, E. 2007. A morphologically driven reference semantic lexicon for French. *4th Workshop on Generative Approaches to the Lexicon*, Paris.

Namer, F. to appear. *Morphologie, Lexique et TAL : l'analyseur DériF*. London: Hermès Sciences Pub.

Pustejovsky J., Boguraev B. (1993), Lexical Knowledge Representation and Natural Language Processing, Artificial Intelligence 63, 193--223.

Pustejovsky J. (1995), The Generative Lexicon, The MIT Press, Cambridge, MA.

Pustejovsky, J. 2001. Type Construction and the Logic of Concepts. Bouillon P. & F. Busa (Eds.) *The Language of Word Meaning*. Cambridge, MA : Cambridge University Press, 91-123.

Ruimy, N. et al. 2003. A computational semantic lexicon of Italian: SIMPLE . Zampolli A., N. Calzolari & L. Cignoni (Eds.) *Computational Linguistics in Pisa. Linguistica Computazionale*, Special Issue, XVIII-XIX. Pisa-Roma : IEPI. Tomo II, 821-864.

Ruimy, N. et al. 2002. « CLIPS, A Multil-level Italian Computational Lexicon: a Glimpse to Data », in : *Proceedings of the 3th Language Resources and Evaluation Conference, Las Palmas de Gran Canaria, mai 2002,* vol. III, éd. The European Language Resources Association (ELRA). Paris, 792-79.

Ruimy N., Bouillon P., Cartoni B. (2005), Inferring a Semantically Annotated Generative French Lexicon from an Italian Lexical Resource. In P. Bouillon, K. Kanzaki (eds.), *GL 2005*, Third International Workshop on Generative Approaches to the Lexicon, Geneva. 218-226.

Scalise, S. 1984. *Generative Morphology*: Studies in Generative Grammar 18. Dordrecht: Foris.