HAL
open science

Extending the Adverbial Coverage of a French WordNet

Benoît Sagot, Karen Fort, Fabienne Venant

To cite this version:

Benoît Sagot, Karen Fort, Fabienne Venant. Extending the Adverbial Coverage of a French WordNet. NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources, May 2009, Odense, Denmark. pp.0. hal-00402305

HAL Id: hal-00402305

https://hal.science/hal-00402305

Submitted on 7 Jul 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.

Extending the adverbial coverage of a French wordnet

Benoît Sagot
INRIA Paris-Rocquencourt / Paris 7
Paris, France
benoit.sagot@inria.fr

Karën Fort
INIST
Nancy, France
karen.fort@inist.fr

Fabienne Venant
INRIA Nancy Grand-Est
Nancy, France
venantfa@loria.fr

Abstract

This paper presents a work on extending the adverbial entries of WOLF, a semantic lexical resource for French. This work is based on the exploitation of the derivation and synonymy relations; the latter are extracted from the DicoSyn synonyms database. The resulting semantic resource, which is freely available, is manually evaluated and validated in an exhaustive manner.

1 Introduction

Nowadays, the availability of resources for Natural Language Processing (NLP) remains a hot topic, in particular for French. The situation is slightly improving as compared to English as far as morphological and syntactic resources are concerned (Sagot et al., 2006). However, this is not yet the case for semantic resources, despite efforts made to provide a freely-available wordnet for French, WOLF (see Section 2.2).

In this paper, we describe a first step in this direction. Restricting our area of investigation to adverbs, our goal is to complete WOLF, thanks to the morphological and syntactic lexicon Le*fff* (Sagot et al., 2006) and the synonyms database DicoSyn (Ploux and Victorri, 1998).

This paper is organized as follows. In Section 2, we introduce the three resources used in our work. In Section 3 we describe how we extended WOLF thanks to two complementary techniques. Finally, in Section 4 we detail the results of the exhaustive manual evaluation of the resulting entries.

2 Ressources

2.1 Le*fff* and the Lexique-Grammaire tables

Le*fff* (Lexique des Formes Fléchies du Français, *Lexicon of French Inflected Forms*) (Sagot et al., 2006), is a large-coverage morphological and syntactic lexicon for French which is freely available.[1] Le*fff* aims at conciliating linguistic relevance and usability in NLP applications. In particular, it is used in several parsers that rely on various formalisms (LFG, TAG). Le*fff*, currently in version 3, covers all categories and is progressively enriched with syntactic and semantic information, notably by comparing it to other syntactic resources (Danlos and Sagot, 2007). Thus, adverbial entries in Le*fff* were enhanced (Sagot and Fort, 2007) thanks to the Lexique-Grammaire tables of adverbs in *-ment*, the so-called Molinier tables (Molinier and Levrier, 2000).

In French, adverbs ending in *-ment* form a large class of adverbs. Moreover, as opposed to other adverbs, it is an open class. Those adverbs form a morphologically homogeneous class, since most of them are built according to the pattern adjective + *ment*. Numerous other adverbs exist, and in particular a large amount of adverbial phrases, but they lie beyond the scope of this work.

2.2 WOLF

WOLF (WOrdnet Libre du Français, *Free French Wordnet*) is a semantic lexical resource for French, freely available (Sagot and Fišer, 2008).[2] It is a *wordnet*, based on the model of the Princeton WordNet (PWN), the first wordnet ever developed, which deals with English (Fellbaum, 1998). Like any wordnet, WOLF is a lexical database in which words (lexemes, literals) are divided by parts-of-speech and organized into a hierarchy of nodes. Each node has a unique id, and represents a *concept* or *synset* (set of synonyms). It groups a certain amount of synonymous lexemes that denote this concept. For example, in the PWN (version 2.0), the synset ENG20-02853224-n contains the

[1] http://gforge.inria.fr/projects/alexina/

[2] http://wolf.gforge.inria.fr/

lexemes {*car, auto, automobile, machine, motorcar*}. Lexemes can be single words as well as multi-word expressions, taking also into account metaphoric and idiomatic usage. Synsets also contain a short gloss, and are related to other synsets. For example, the above-mentioned synset is related to the synset {*motor vehicle, automotive vehicle*} by a hypernymy relation, and to the synset {*cab, hack, taxi, taxicab*} by a hyponymy relation.

WOLF was built using the PWN 2.0 and various multilingual resources, thanks to two complementary approaches. Polysemous lexemes were dealt with using an approach that relies on parallel corpora in five languages, including French, that were word-aligned. Several multilingual lexicons were extracted from those aligned corpora, taking into account three to five of the available languages (precision and recall of these lexicons vary w.r.t. the number of languages taken into account). Multilingual lexicons were semantically disambiguated thanks to wordnets for the corresponding languages. On the other hand, monosemous PWN lexemes only required bilingual lexicons that were extracted from wiki resources (Wikipedia, Wiktionary) and thesauri. Nominal and verbal sub-wordnets of WOLF were evaluated against the French wordnet built during the EuroWordNet project.[3]

WOLF contains all PWN 2.0 synsets, including those for which no French lexeme is known. The latest version of WOLF before this work, version 0.1.4, includes French adverbial lexemes for only 676 of the 3,664 adverbial synsets, i.e., only 18.4%, and only 983 lexeme-synset pairs corresponding to only 665 unique adverbial lemmas. For this reason, we applied two complementary techniques to improve WOLF's coverage. One of those techniques relies on the morphological and semantic derivation relation that often exists between an adverbial synset and its corresponding adjectival synset, both in English and French. The other technique relies on the expoitation of the synonyms database DicoSyn.

[3]The wordnet developed during the EuroWordNet project (Vossen, P., 1999) is the only other French wordnet. It contains only nominal and verbal synset, but no adjectival or adverbial synsets. Moreover, important license problems explain why it is rarely used in the research community. Finally, and partly for the same reason, it has not been improved since its creation. Those are the three main motivations for the development of WOLF.

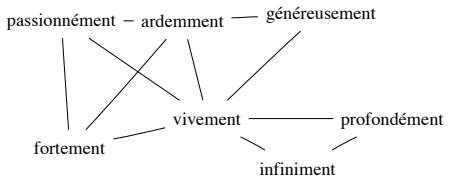

Figure 1: Extract from the adverbial synonymy graph

2.3 DicoSyn and the cliques of synonyms

DicoSyn is an electronic dictionary of synonyms, whose latest versions are available for online usage.[4] The initial base (Ploux and Victorri, 1998) was created merging seven French classic dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert) from which the synonymic relations were extracted. The major advantage of this dictionary is that it explicitly shows the graph of the synonymy relation.[5] Ploux and Victorri designed Visusyn, that allows to explore the graph. It is then possible to automatically visualize and characterize the semantic properties of a unit, using the sub-graph it constitutes with its synonyms (François et al., 2002; Venant, 2004), or to study in a more global way the semantic characteristics of a whole lexical paradigm (Venant, 2007). We were thus able to exploit a graph of adverbial synonyms. As DicoSyn does not contain any indication regarding categories, this graph was built mapping DicoSyn with the adverbs in *-ment* from Le*fff*. The graph comprises 1,597 nodes (adverbs) and 4,344 (synonymy) connections. Among those nodes some are not adverbs ending in *-ment*, but synonyms of such adverbs (for example, *bien* is a node of the graph due to the fact that DicoSyn indicates that it is a synonym of *amplement* or *copieusement*). Figure 1 presents an extract from this graph.

We exploited this graph using the notion of clique. A clique is a the largest possible set of nodes connected as pairs. Thus, the graph in figure 1 contains 3 cliques: {*ardemment, fortement, passionnément, vivement*} (we cannot add *généreusement* which is neither a synonym

[4]http://dico.isc.cnrs.fr/ and http://elsap1.unicaen.fr/dicosyn.html

[5]It is of course a partial synonymy relation, that Ploux and Victorri define in the following way: "Two lexical units are in a synonymy relation if any of the occurrences of one of them can be replaced by any occurrence of the other in a certain number of environments, without profoundly modifying the meaning of the text it belongs to." It is a symmetric relation.

of *fortement*, or *passionnément*), {*ardemment, généreusement, vivement*} and {*infiniment, profondément, vivement*}. The obtained adverbial graph comprises 2,247 cliques. The idea behind this is that a clique corresponds to a possible usage of the adverb. A clique being a set of synonyms, it more or less corresponds to a WordNet synset. Thus, cliques constitute the structural unit of the graph semantic analysis.

3 Extending WOLF

As previously stated, we first extended WOLF in order to increase the number of non-empty adverbial synsets (for which at least one French lexeme exists) as well as the number of lexemes in each non-empty synset. To do so, we used two types of relations between lexemes: the derivation relation, between an adverb ending in *-ment* and its corresponding adjective, and the synonymy relation between adverbs, as defined by the cliques in DicoSyn .

3.1 Using the derivation relation

The method based on the derivation relation arose from the two following observations:

- The PWN includes a derivation relation (*derived*) that links some adverbial synsets to one or more adjectival synsets. This link indicates that some adjectival lexemes in the adjectival synset allow the construction, using morphological derivation (*-ly* suffix), of some adverbial lexemes of the adverbial synset. Naturally, this link also indicates a semantic connection between the two synsets.

- The mechanism of morphological and semantic derivation between adjectives and adverbs is often parallel in English (adjective + *ly*) and French ($\text{adjective}_{\text{fem,sing}}$ + *ment*).[6]

We therefore collected, for each adverbial synset, the (French) adjectives in the adjectival synset connected through the *derived* relation. We then applied the morphological derivation algorithm to those adjectives.[7] The obtained adverbs which appear in Le*fff* were kept and allocated to the adverbial synset (with a note specifying that the lexeme–synset links were built using morphological derivation).

Let us consider, for example, the ENG20-00115661-b synset. In WOLF 0.1.4, it only contains the (correct) lexemes *toujours* and *invariablement*. Yet, this synset is connected to the adjectival synset ENG20-02417249-a through a *derived* relation and the latter comprises the lexemes *permanent*, *invariable* and *perpétuel*. Therefore, the potential adverbs *permanentement*, *invariablement* and *perpétuellement* are built. The first one is removed, as it does not appear in Le*fff*, the second one confirms a lexeme that already belonged to the adverbial synset, and the last one allows the creation of a new lexeme–synset connection. In the end, the ENG20-00115661-b synset is transformed into {*toujours*, *invariablement*, *perpétuellement*}.

Using this method, the number of adverbial lexeme–synset relations in WOLF raised from 983 to 1,536 (+56%). The number of non-empty adverbial synsets raised from 676 to 969 (+43%). The number of adverbial lexemes in WOLF raised from 665 to 889 (+23%).

3.2 Using the synonymy relation

Once the adverbial synsets of WOLF completed using the derivation relation between adverbs ending in *-ment* and adjectives, we applied a method based on the synonymy relation, as defined by the DicoSyn cliques. Three steps were necessary.

1. We first associated to each lexeme–synset connection a weighting rate according to their origin (see section 2.2). If a connection was built (among other sources) from bilingual resources (wiki resources), it receives a rate of 5. If the connection was built using aligned multilingual corpora, the rate is 4, if one of the corpus contained at least 4 languages, 3 if they all contained only 3 languages. In all other cases, including for connections built using the derivation relation, a rate of 2 is associated to the connection.

2. Each adverbial synset is then associated to the DicoSyn clique which corresponds the most, i.e. not simply containing the highest number of lexemes in common, but rather maximizing the sum of the rates of the lexemes shared by the clique and the synset.

[6] This is of course not always true (see courante/couramment and many others), but it is still a reasonable heuristics.

[7] The feminine singular form of the adjective being taken from Le*fff*.

3. Each synset is then completed with all the lexemes (adverbs) belonging to the associated clique.

For example, let us consider the ENG20-00115661-b synset, the very same synset we previously detailled. Once extended using the derivation relation, it contained the adverbs *toujours*, *invariablement* and *perpétuellement*. As the first two were built using the French Wiktionary, they receive a rate of 5. The adverb *perpétuellement*, built by derivation, receives a rate of 2. Therefore, the clique maximizing the sum of the rates of the common lexemes is {*éternellement*, *invariablement*, *perpétuellement*, *sans cesse*, *toujours*}. Two adverbs were thus added to the ENG20-00115661-b synset, the multi-word adverb *sans cesse* and the *-ment* adverb *perpétuellement*.

Using those methods, we increased the number of lexeme–adverbial synset relations from 1,536 to 2,149, which represents a 28.5% increase.

4 Evaluation of the extended WOLF

4.1 Methodology

We conducted a manual evaluation of all the adverbial synsets we obtained, i.e. of the 2,149 lexeme–synset pairs, comprising 1,025 adverbial lexemes. Each author manually validated the couples comprising one fourth of the lexemes; the remaining fourth being evaluated by the three authors, thus allowing for inter-validator agreement calculus.

Validating a lexeme–synset pair consists in assigning it one of the following codes:

- OK: correct association;
- SC (Semantically close): one of the meaning of the lexeme is semantically close to that of the synset (hyponym, hypernym, pseudo-synonym);
- SR (Semantically related): one of the meaning of the lexeme is semantically related (but less close) to that of the synset;
- NR (Non Related): no meaning of the lexeme is related to that of the synset;
- CC (Composed Component): false association, but the lexeme is one of the component of a multi-word lexeme which would fit in the synset;

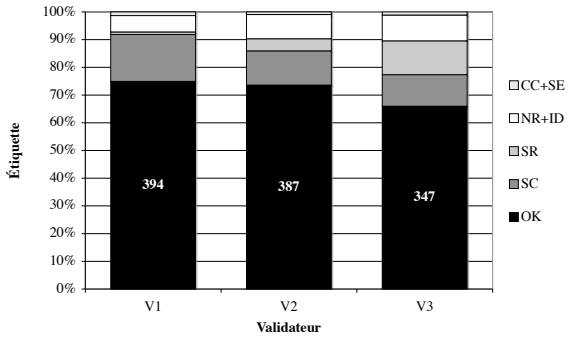

Figure 2: Comparison between the 3 validators of the distribution of evaluation codes for the same 530 lexeme–synset pairs

- ID (Incorrect Derivation): false association, due to a derivation issue such as an ambiguity of the intermediary adjective or the lack of parallel between morphological and semantic derivation (see, for example, *absolument* in the synset defined by *in a royal manner*)
- SE (Spelling Error): spelling error in the lexeme, the association is to be rejected;
- WC (Wrong Category): false association, due to an erroneous part-of-speech tagging of the lexeme (see, for example, *bougonnerie*)

4.2 Inter-validator agreement

For one fourth of the lexemes, the three authors carried out the evaluation independently. If we replace all the codes other than OK by a unique NONOK code, the three validators agree on 366 of the 530 lexeme-synset pairs, i.e., 69% of such pairs are validated three times NONOK or three times OK. The latter case (all validators agree the pair is correct) covers 292 lexeme-synset pairs (55%). Examining the distribution of the codes for each validator, we noticed differences in terms of tolerance level (see figure 2). As the boundary between codes like SC, SR and NR is difficult to define objectively, the variety of decisions about them is not surprising. On the opposite, over the 456 pairs judged OK by at least one of the validators, only 292 were validated (OK) by the three validators (64%) and 94 by two validators (20,6%). The agreement rate is therefore quite low. This can be explained by the difficulty of the task (some synsets cannot be easily differentiated) and by the scarcity of some adverbs.

The analysis of those results led us to associate a unique code to the lexeme–synset pairs evaluated by the three validators, in the following way:

- OK if the three evaluations are OK-OK-OK, OK-OK-SC, OK-OK-SR, or OK-SC-SC ;
- SC if they are OK-SC-SR or SC-SC-SR ;
- SR in the other cases where there is one or two OK amongst the three, as well as in the SC-SR-SR and SR-SR-SR cases;
- SE (ID, CC, WC) in the other cases, if a validator gave the SE code (ID, CC, WC);
- NR in the remaining cases.

Needless to say that the lexeme–synset pairs evaluated by only one validator keeps the code s/he gave them.

4.3 Evaluation results and obtained resource

The results are quite promising (see table 1), as we obtain more than 68% of correct lexeme-synset associations (OK). We kept 1,461 of the 2,149 lexeme-synset relations that we built automatically (as compared to 983 before this work, which were not manually validated). WOLF now contains 871 adverbial lexemes (as compared to 665 when we started) belonging to 871 non empty synsets (as compared to the initial 676). Therefore, the improvements in WOLF cover not only its quality, due to the manual validation, but also the number of synsets.

Total	OK	SC	SR	NR
2 145	**1 461**	296	147	162
100%	**68,1%**	13,8%	6,9%	7,6%

ID	CC	WC	SE
41	26	13	3
1,9%	1,2%	0,6%	0,1%

Table 1: Results of the manual validation

5 Conclusion and prospects

At a time when the lack of large scale lexical resources for French weights on NLP research, we showed the interest of using several existing resources to enrich or diversify their content. The Le*fff*–WOLF interaction, through DicoSyn, allowed us to enrich WOLF both in terms of quality and quantity. This work led to an increase of nearly 55% of the adverbial lexeme–synset relations in WOLF.

Those encouraging results also show that it is worthwhile exploiting a lexicon as a graph, at least as far as the automatic access to semantic information is concerned. The synonymy and the adverbs ending in *-ment* were ideal for this experiment and encourage us to explore other paradigmatic (hypernymy, antonymy) or syntagmatic (through corpus analysis) relations, as well as other parts-of-speech, like, for example, the nouns ending in *-ité* or the verbs in *-ifier* and *-iser*.

References

Laurence Danlos and Benoît Sagot. 2007. Comparaison du Lexique-Grammaire et de Dicovalence: vers une intégration dans le Le*fff*. In *Actes de TALN 07*, Toulouse, France.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jacques François, Bernard Victorri, and Jean-Luc Manguin. 2002. Polysémie adjectivale et synonymie : l'éventail des sens de curieux. *La polysémie*.

C. Molinier and F. Levrier. 2000. *Grammaire des adverbes. Description des formes en* -ment. Droz, Geneva, Switzerland.

Sabine Ploux and Bernard Victorri. 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues (T.A.L.)*, 39(1):161–182.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Actes de Ontolex 2008*, Marrakech, Morocco. (à paraître).

Benoît Sagot and Karën Fort. 2007. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – adverbes en -ment. In *Actes du Colloque Lexique et Grammaire*, Bonifacio, France.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use. In *Proc. of LREC'06*.

Fabienne Venant. 2004. Polysémie et calcul du sens. In *Actes de JADT 2004*, Leuven, Belgium.

Fabienne Venant. 2007. Une exploration géométrique de la structure sémantique du lexique adjectival franais. *Traitement Automatique des Langues (T.A.L.)*, 47(2).

Vossen, P. 1999. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.