Acquiring semantics from structured corpora to enrich an existing lexicon

Núria Gala, Véronique Rey

Acquiring semantics from structured corpora to enrich an existing lexicon

Nuria Gala[1], Véronique Rey[2]
Aix-Marseille Universités, France

Abstract

Lexical and semantic information is capital in linguistic resources, be it for language learning purposes or for NLP applications. However, this kind of information is very difficult to collect either manually or automatically. The difficulties come from the nature of the information (what do we mean by 'semantics'? how is the 'meaning' put into words?) and from the resources themselves (how are the 'semantics' formalised and displayed?). In this paper we present a methodology to automatically acquire semantics from structured corpora (machine-readable dictionaries and free encyclopaedias on the web). This information is used to enrich an existing lexical database for constructional families of words in modern French.

Keywords: automatic acquisition, semantic information, morpho-phonology, constructional families, lexical database.

1. Introduction

In the Age of generalized electronic linguistic resources, language learners benefit from a variety of available applications. As for monolingual dictionaries or lexical databases are concerned, given some word, the learner may look up its meaning, some etymological, phonological and grammatical information, as well as several examples illustrating the different usages in which the word may be produced. In addition, the notion of how common a word is, introduced during the nineties by means of frequencies extracted from corpora (Kilgarriff, 1996), is now significantly wide-spread. The learner may thus acknowledge if the word is common or less common in the language or in a particular specialized sub-language.

The resources used for e-learning tend to be electronic versions of paper resources: they present the same information, the only difference being the way this information is searched and displayed. Some examples illustrate this point for French: the TLFi[3] (*Trésor de la Langue Française*) and the DAF[4] (*Dictionnaire de l'Académie*

[1] LIF, CNRS UMR 6166, Marseille, France, nuria.gala@lif.univ-mrs.fr
[2] SHADYC, CNRS & EHESS, UMR 8562, Marseille, veronique.rey@univ-provence.fr
[3] http://atilf.atilf.fr/tlf.htm
[4] http://dic.academic.ru/dic.nsf/daf_1835

Française), both available online with exactly the same content as in the paper version.

Lexical databases and networks, built to be used by means of computers (no paper versions), are supposed to go further on the lexical description (due to storage possibilities, to the conceptual organisation of the information, etc.). Though a number of projects have arisen, specially for multilingual resources (Papillon database[5], EuroWordNet, etc.), the learner only obtains definitions and lists of synonyms when looking for semantic information. It is not feasible for such resources to "navigate" through lexical units sharing semantic components in a same constructional family neither to access a particular lexical unit from a set of ideas carried by the semantic information (as in flexional languages like French the construction of words is based on phonological and semantic continuity, it would be relevant to navigate in a resource by exploiting this implicit practice of speakers).

The objective of the work described in this paper is to propose a method for automatically enrich with semantic information an existing lexical database for modern French, POLYMOTS. Yet enriching this resource with semantics may allow a variety of functionalities for navigation and lexical access.

The paper is organized as follows: first, in section 2, we introduce the notion of morpho-phonological families and we briefly outline the main features of POLYMOTS. Along section 3 we present a method for automatically collect semantic information from structured corpora and we discuss on the notion of continuity and dispersion of meaning within a family of words. Finally, we present our conclusions and future work in section 4.

2. Morpho-phonological families in modern French

Traditionally, lexical morphology has been diachronic and has focused on the notion of word families on the basis of word form origins (etymology). In morphological synchronic studies, the focus is rather on segmenting words in minimal meaningful units (morphemes). In this approach, the aim is to build models of morphological constructions of the lexicon. However, this task is far from being trivial.

1.1. Morpho-phonology

If in some cases structural analysis is less complex, *i.e.* "bras" (*arm*) and "brassard" (*armband*) clearly share a common stem "bras" (*arm*, which is also common in both English translations), in some other cases, segmentation is not straightforward because a common unit is more difficult to seize. The following are two examples of questions that may arise: (1) do "biscuit" (*cookie*) and "cuire" (*to bake*) belong to the same family? ; (2) can we regroup into a same family "*confiture*" (jam) and "défaite" (*defeat*)? In the first example, it is possible to recognize the past participle form of the

[5] http://www.papillon-dictionary.org

French verb to bake ("cuit", *baked*) ; in the second one, the simple past form ("fit", *did*) and the past participle ("fait", *done*) of the same verb "faire" (*to do*). In both cases, we are not identifying exactly the same form along the members of a family but rather *one of the possible forms*.

As a consequence, and following Kiparsky (1982), we assume that the process of word construction implies phonological transformations, that is, vocalic and consonantic alternations. As the process takes place within the lexeme, we can talk about morpho-phonological processes. However, the alternations are not systematic; to give an example, many words ending in /o/ alternate with /el/ ("ciseau" *chisel*, "ciseler" *to chisel*; "château" *castle*, "châtelain" *manor*; "appeau" *decoy*, "appel" *call*) but others do not ("fourreau" *sleeve*, "berceau" *cradle*, "gâteau" *cake*). To take into consideration this morphophonemic process, the words in our word family corpus have been manually annotated.

1.2. Semantics

Segmenting words into morphemes raises interesting questions about meaning, as regards to morphemes themselves and to word families as a whole.

1.2.1. Morphemes

Some words in French have been created with meaningless morphemes (we call them *opaque stems*). To give an example, the word "tri-maran" has been built following "cata-maran", although "maran" is a non latin meaningless stem. Another example is that of "panta-lon" and "panta-court" which share the stem "panta".

In other cases, it is possible to identify a common stem in a family which has not a meaning as a single word in modern French (although a latin origin). For example, "duct" is not a word as a single unit; however, it is present in "con-duct-eur" (*driver*), "pro-duct-eur" (*producer*), "intro-duct-ion", "sé-duct-eur" (*seducer*), etc.

1.2.2. Word Families

The point that we want to highlight in this paper is mainly concerned with meaning within word families. As explained in section 1.1., words are grouped into families on the basis of a common morpho-phonological stem. Therefore, in some families the segmentation entails a question of lexical variation. To illustrate this idea, if words in the previous family share the same stem ("duct"), the question that raises at this point may be: what is the semantic link between all the words in this family ? This question comes from our hypothesis that all the words in a family share not only a morpho-phonological stem, but also a semantic coherence. In some cases, the common meaning appears quite straightforward: "terre" (*globe/earth*), "territoire" (*territory*) and "terrasse" (*terrace*) share the notion of *area* and *surface* ; "gluant" (*viscous*), "glutineux" (*sticky*), "agglutiner" (*aglutinate*) may share the notions of *sticky, adhesive, viscous*, etc.

Clearly, the semantic degree of cohesion in a family is very different. In some cases it is transparent, in other cases it is more difficult to grasp and, as a consequence, it entails substantial conjectures about the semantic continuum in a family. That is the point we wanted to investigate by automatically acquiring semantic information from structured corpora (cf. Section 2). Before describing our method, we briefly outline the features of the lexical database developed for French words segmented into families.

1.3. POLYMOTS

As a result of manual segmentation of 20.000 French words, we have obtained a lexical database containing about 2.000 families (Gala & Rey, 2008). The morphological analysis has allowed us to identify about 1/3 *opaque stems* and 2/3 *transparent stems* (in this case, the stem is a meaningful French word, for example: “terre” (*earth/globe*), “glue” (*glu*), “boule” (*ball*), etc.).

As constructional morphology is very common in French -and other Romance languages-, the average of words in a family is about ten lexical items. However, productivity is very different within the families. Unlike families with only one or two members, for example “chaise” (*chair*) constitutes the only lexical unit of its family, “choi” is the common stem in “choix” (*choice*) and “choisir” (*to choose*), some stems can be found in families containing up to seventy or eighty lexical items (“mue/mut” in “commuter” (*to commute*), “immuable” (*immutable*), “mutuel” (*mutual*), “remuer” (*to shake*), etc.).

3. Semantic information acquisition from structured corpora

At present, as a lexical database POLYMOTS only displays morphological information and is being used by speech therapists for improving the vocabulary learning task of patients presenting particular diseases (dyslexia, Alzheimer). The need of semantic information in this context is twofold: to understand unknown lexical units and access them.

We thus consider that the learner would better understand the meaning of a lexical unit by grasping the semantic links with other words belonging to the same family (given some unknown words as *gluey* or *glueball* s/he would be able to catch their meanings by comparing it to *glue*, the ‘baseword’, which shares with the former unknown words the notions of *sticky, adhesive, viscous*). Following (Zock & Schwab, 2008), adding semantics will also allow the learner in another perspective: s/he will be able to find a precise lexical unit from a set of ideas (taking *file, key, path, fast* s/he will accede to *shortcut*).

3.1. Related work on semantic acquisition

Automatically acquiring information from available corpora is one of the classical tasks in natural language processing (NLP). However, the construction and enrichment

of electronic resources from corpora is far from being trivial, particularly due to the availability of data. Let aside manually built resources (extremely time-consuming), a number of works have been done using the web (Grefenstette 2007).

Other approaches have been using existing resources such as dictionaries, synonym lists, ontologies, etc. because the information encoded is completely structured and thus easily available. The use of dictionary definitions for different applications is thus widespread: to create lexical networks (Ide and Véronis 1990), to build an example database used for semantic disambiguation (Brun et al. 2001), for the acquisition of conceptual links between words (L'Homme 2003), to build lexical graphs (Gaume et al. 2007), etc.

3.2. Structured corpora

Focusing on our word families, the use of structured corpora is crucial for enriching every word in a family. The idea here is to collect information from different structured resources to obtain a list of semantic units describing each word.

One of the difficulties in using structured corpora is their availability. For reasons of copyright, most of the dictionaries with online consultation are not available in an exploitable format (text format or, better, XML format). That is the case of one of the main French dictionaries, the TLFi (*Trésor de la Langue Française Informatisé*).

As we wanted to diversify our sources, we thus used the following lexicographic an encyclopaedic resources:

- Hachette Multimédia dictionary (in XML format)
- Wiktionnaire[6] (French version of Wiktionary)
- French Wikipedia[7]

Our aim was to collect meaningful words present on the definitions of the dictionaries and in the introduction paragraph of Wikipedia. Using the list of 20 000 words we collected the different entries of those sources.

In the case of Wikis, we retrieved the required web pages, corresponding to our list of words (we used Lynx, a Linux text-only web browser).

From Wikipedia, we only retrieved the introductory paragraph of each article, in order to avoid 'noisy' with encyclopaedic concerns (which we considered less relevant for our purpose). After removing the HTML tags, we obtained text files as shown on the following figure for the word “vache” (cow) used to illustrate our method along this section:

[6]http://fr.wiktionary.org/wiki/

[7]http://fr.wikipedia.org/wiki/

vache féminin (Zoologie) Mammifère domestique ruminant, généralement porteur de cornes sur le front, appartenant à l'espèce *Bos taurus* de la famille des bovidés. Femelle de cette espèce.

Figure 1. Sample obtained from Wiktionnaire.

Vache (Brune Suisse ou Brune des Alpes) vue sous la Fuorcla Sesvenna dans l'Engadine, en Suisse. La vache est la femelle d'un mammifère domestique ruminant, généralement porteur de cornes sur le front, appartenant à l'espèce Bos taurus de la famille des bovidés. C'est la femelle du taureau. Une génisse est une vache qui n'a pas vêlé. Le poids moyen d'une vache adulte varie en fonction de la race de 500 à 900 kg. Le mot vache vient probablement du sanscrit Vaça désignant une génisse qui vêle pour la première fois.

Figure 2. Sample obtained from Wikipédia.

3.3. Methodology to obtain semantic units

We tested the hypothesis that each definition would contain significant lexical items to semantically characterize the words from our families. We thus regrouped the corpora by headword and extracted the meaningful words (we removed stopwords such as prepositions, articles, some frequent adverbs, conjunctions and a number of lexicographic nouns such as 'verb', 'synonym', 'example', 'latin', etc.). As we were interested in lemmas, in order to count as a single item any flexional variation of a word, we used Treetagger[8], an available part-of-speech tagger and lemmatizer.

For each entry, the Treetagger output was transformed into a vector of words as the following:

vache	femelle bovin
vache	manoeuvrer attaque sournois
vache	peau cuir animal
vache	récipient plier toile plastique analogue utiliser campeur

Figure 3. Meaningful lemmatized words from Hachette definitions.

Notice that each definition is kept with its headword, because the information about the order of the lexical elements is significant for calculating the importance of a meaningful word.

[8]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

The impact of errors from Treetagger, specially in ambiguities noun/verb remains a weakness of the approach. In the following examples, nouns have been tagged as verbs: “manoeuvrer” (*to maneuver*), instead of “manoeuvre” (*maneuver*); “corner” (*to corner, to honk*) instead of “corne” (*horn*); “membrer” (*~to equip*) instead of “membre” (*member*), etc. There is work in progress to fix this problem by means or further tagger training.

3.4. Weighting semantic units

Once the meaningful words, called hereafter semantic units, have been lemmatized, our aim is to evaluate their “importance”, that is, empirically calculate their relation with the headword of the definition. As our corpora is made of three different sources, we made the hypothesis that significant units will appear in most of the definitions and generally at the beginning of them. To give an example, for “vache” (*cow*), the word “femelle” (*female*) is more significant than “génisse” (*heifer*): the first one appears twice (at the beginning of two definitions), the later in the middle of a single definition (to be precise, an encyclopaedic sentence). The relevance of those words has to be taken into account differently.

We have thus attributed a wheight to each word within a definition, taking into account the distance from each word the headword. To put it in formal notation, for heach headword w we have a list $\alpha(w)$ containing semantic units u. Given a semantic unit u_i present within this list, we calculate the weight $\omega(u_i)$ according to the distance of u_i to w and taking into account the total number of words n in each definition (with $0 \leq i < n$):

$$\forall\, u_i \in \alpha(w),\ \omega(u_i) = 1 - i/n$$

The relevance of a semantic unit u_i decreases taking into account its distance to the headword w. If there are four words in the definition, the first one will be assigned 1, the second $1 - 2 / 4 = 0.5$, the third $1 - 3 / 4 = 0.75$ etc. (Gala et al. 2009).

A final adjustment is necessary for words appearing more than once within the definitions for a same headword. In this case we add theirs weights and harmonize them with the addition of all weights (bringing them to a maximal $\omega = 1$ and always $\omega > 0$). To give an example, *femelle* ('female') appears twice with $\omega = 1$; *mammifère* ('mammal') appears twice with $\omega = 0.94$ and $\omega = 0.93$. In this case, the final weight will be 0.58. The following figure shows the final results for the entry *vache* ('cow').

[femelle 1.00] [mammifère 0.58] [domestique 0.54] [ruminer 0.50] [porteur 0.45] [espèce 0.43] [corner 0.41] [front 0.37] [appartenir 0.32] [adulte 0.31] [manoeuvrer 0.31] [peau 0.31] [récipient 0.31] [vêler 0.31] [zoologie 0.31] [plier 0.27] [varier 0.25] [bos 0.23] [toile 0.22] [attaque 0.21] [cuir 0.21] [poids 0.21] [taurus 0.19] [fonction 0.19] [plastique 0.18] [bovin 0.16] [famille 0.15] [analogue 0.13] [race 0.13] [animal

0.10] [moyen 0.10] [sournois 0.10] [bovidés 0.10] [utiliser 0.09] [mot 0.06] [campeur 0.04] [taureau 0.04] [génisse 0.02]

Figure 4. Semantic units obtained after weighting

As cow is a polysemic word in French, the semantic units within the vector show such a variety of meanings: "mammifère" *(mammal)*, "cuir" *(cowhide)*, "sournois" *(rotten, mean)*, "toile" *(tent).* A vector obtained with this method may contain synonyms (*embrace* and *enclose*, *swallow* and *go down*), hyperonyms (*mammal* and *heifer*, *alarm* and *device*) as well as thematic links (*alarm* and *enemy*, *swallow* and *throat*), etc.

4. Continuity *vs* dispersion of meaning in a family

Morpho-phonological families are based on two methodological criteria: phonology and semantics. As for semantics, a thorough study of the semantic units present within the vectors has lead us to identify the following two concepts.

4.1. Semantic continuity

Semantic continuity is the property of families sharing a number of semantic units. To be precise, some semantic units are kept within the family and, in most cases, a recurrent word (the transparent stem) is present in the vectors of all the family members. The following figure illustrates some of such families:

terre (*earth/globe*)	surface	bras (*arm*)	membre (*member*)
territoire (*territory*)	surface	brassard (*armband*)	bras (*arm*)
terrasse (*terrace*)	surface	embrasser (*to embrace*)	bras (*arm*)
		bracelet (*bangle*)	bras (*arm*)

Figure 5. Semantic continuity

There is an explicit continuity of meaning among the words carried out by a same semantic unit throughout the family.

4.1. Semantic dispersion

Semantic dispersion is the property of families where a common semantic unit is not present throughout *all* the words in the family. In these cases, only one, or few semantic units are shared (between a word in the family and the 'headword' or stem).

fil (thr*ead*)	long, continuité, fin (*long, continuity, thin*)
défilé (*parade*)	long, continuité
profil (*profile*)	fin
val (*glen*)	aire, descente (*area, downhill*)
vallée (*valley*)	aire

avaler (*swallow*)	descente

Figure 6. Semantic dispersion

Recurrent semantic units characterize the words in families with dispersed meaning. Even if these meaningful units may be different, they are all to be found within the vector characterizing the 'headword' corresponding to the stem. In cases where the stem corresponds to a non existing word in modern French (*opaque stems, cf.* Section 1.2.1), a number of common semantic units are to be found within the family, i.e. the notion of *dead* and *dangerous* within the members of the family sharing the opaque stem “cid” (“accident”, “suicide”, “incident”, “acide”, etc.).

5. Conclusion and future work

In this paper, we have presented a method for automatically enrich an existing lexical database for French with semantic information. The initial project was to create a lexical resource of word families based on word constructions and following a morphophonemic approach. As we assumed that words within the families shared common semantic features, we gathered the semantic information from structured corpora to empirically validate our hypothesis. Our first results entail two types of families, depending on the dispersion or the continuity of meaning between the words in a family.

Word families in POLYMOTS offer a new perspective on the study of words based on phonological stems resulting from language usages, instead of traditional lemmas anchored in diachrony. This resource is thus an example of a new approach in e-lexicography offering different possibilities to learn French vocabulary on a basis of phonology and semantics.

References

BRUN C., JACQUEMIN B., SEGOND F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique. *Revue TAL, Traitement Automatique des Langues,* volume 42(3), pages 667 à 691.

GALA N., REY V. & TICHIT, L. (2009) Dispersion sémantique dans des familles morpho-phonologiques: éléments théoriques et empririques. *Actes de TALN 09: Traitement Automatique des Langues Naturelles*, Senlis, Juin 2009.

GALA N. & REY V. (2008) POLYMOTS : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. *Actes de TALN 08: Traitement Automatique des Langues Naturelles,* Avignon, juin 2008.

GAUME B., DUVIGNAU K., VANHOVE M. (2007) *Semantic associations and confluences in paradigmatic networks*. In: M. Vanhove (éd.), Typologie des rapprochements sémantiques.

GREFENSTETTE, G. (2007). Conquering Language: Using NLP on a Massive Scale to Build High Dimensional Language Models from the Web. *Computational Linguistics and Intelligent Text Processing*. Springer Berlin / Heidelberg, 35-49.

IDE N. ET VÉRONIS J. (1990) Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries, *13th International Conference on Computational Linguistics COLING'90,* Vol. 2, pp. 389-394. Helsinki.

KIPARSKY P. (1982). From cyclic Phonology to lexical Phonology. *The structure of Phonological Representations* (1). V. H and S. N. New York, Foris Dordrecht: 131-175.

L'HOMME M. C. (2003). Acquisition de liens conceptuels entre termes à partir de leur définition. *Cahiers de lexicologie* 83 (2), pp. 19-34.

SCHWAB D., LIAN TZE L., LAFOURCADE M. (2007) Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux. *Actes de TALN 07: Traitement Automatique des Langues Naturelles*, Toulouse, 5-8 juin 2007, pages 293-302.

KILGARRIFF, A. (1996). Putting frequencies in a dictionary. *International Journal of Lexicography*.

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In the *Fifth International Conference on Systems Documentation, ACM SIGDOC*.

ZOCH, M. & SCHWAB, D. (2008). Lexical Access based on Underspecified Input. In *Proc. of COLING Workshop endorsed by SIGLEX, Cognitive Aspects of the Lexicon: Enhancing the Structure, Indexes and Entry Points of Electronic Dictionaries*, Manchester, UK.