From the Definitions of the *Trésor de la Langue Française* to a Semantic Database of the French Language

Lucie Barque (STL, Université Lille 3)

Alexis Nasr (LIF, Université Aix-Marseille 2)

& Alain Polguère (ATILF CNRS, Nancy Université)

1. Introduction

The *Definiens* project aims at building a database of French lexical semantics that is formal and structured enough to allow for a fine-grained semantic access to the French lexicon—for such tasks as automatic extraction and computation. To achieve this in a relatively short time, we process the definitions of the *Trésor de la Langue Française informatisé* (Dendien & Pierrel, 2003; TLFi, 2004)—hereafter TLFi—, enriching TLFi's definitions with an XML tagging that makes explicit their internal organization. Definiens builds on the results of the BDéf project (Altman & Polguère, 2003), that derived a small database of fully formalized definitions from the *Explanatory Combinatorial Dictionary of Contemporary French* (Mel'čuk *et al.*, 1984, 1988, 1992, 1999). There is, to our knowledge, no existing broad coverage database for the French lexicon that offers to researchers and NLP developers a structured decomposition of the meaning of lexical units. Definiens is an ongoing research that will hopefully fill this gap in the near future. In Section 2, we explain the problem at hand. Then, in Section 3, we detail the first stage of definition structuring, namely the segmentation into definitional components, and explain how this operation can be achieved by automatic procedures. In Section 4, we detail the second stage of definition structuring, which consists of enhancing TLFi's definitions with semantic labels.

2. Context and aims of Definiens

2.1 Using analytical definitions

There are two main reasons why we decided to start from the *Trésor de la Langue Française*, rather than from any other dictionary, in order to build a French semantic database. First of all, there is the obvious reason that the TLF was available to us for research purposes in electronic form, as TLFi. All other large-scale descriptions of the French lexicon are commercial products or are embedded in commercial products, and therefore not available for this kind of research. Second, TLF's lexicographers have adopted what we view as good practice in structuring their definitions: most of TLF's definitions belong to a specific class of lexicographic definitions that have been termed **analytic definitions** in Polguère (2008:182-188). Analytic definitions correspond closely to so-called "definitions by genus and diffentiæ", which makes them compatible with the type of formal structuring we are targeting. Analytic definitions possess the two following properties:

1. they are analytic paraphrases of the lexical unit under description—the term *analytic* implies that they decompose the meaning of this unit in terms of simpler meanings (they are not lists of synonyms or quasi-synonyms);
2. they are made up of two main parts: a) the **genus**—or **generic/central component** of the definition—, that stands in a hyperonymic relation with the defined unit, and b) the **differentiæ**—or **peripheral components**—, that specify the unit's meaning relative to the genus and to other semantically related lexical units.

This type of definition is the norm in most reference dictionaries, especially in learners' dictionaries, sometimes with some "pedagogical" adaptation, such as in the Cobuild (Sinclair, 1990; Rundell, 2008). Analytical definitions are also used in theoretical dictionaries, such as Mel'čuk (1984, 1988, 1992, 1999) and Wierzbicka (1987), and have proved very useful for the study of the semantics of natural languages—for instance, in the formal study of French polysemy (Martin, 1979; Barque, 2008).

2.2 Advantage of making explicit the internal structure of definitions

We believe analytical definitions to be the most straightforward and natural way of describing lexical meanings. Though it is Aristotle—cf. *Topics* (Aristotle, 1939 translation)—who is credited for their first theoretical conceptualization, they are probably as old as natural languages themselves as they correspond to the most spontaneous way of accounting for meanings in a metalinguistic fashion (Q: — *What does "X" mean?* A: — *It means ...*). From a natural language processing or, more generally, formal point of view, however, analytical definitions present one major drawback: they are "texts" rather than fully-formalized structures— like typed feature structures proposed by the Generative Lexicon (Pustejovsky 1995), for instance. One way to remedy, at least partially, this problem is to make explicit the internal structure of text definitions by means of an XML tagging that will identify:

1. the central and peripheral components of the definition;
2. the normalized content of the central component—encoded by means of **semantic labels** (Polguère, 2003);
3. the normalized semantic role played by each peripheral component relative to the central component—encoded by means of a small subset of semantic labels.

This tagging will superimpose on the definitional text a fully-formalized semantic skeleton that will allow for non-trivial automatic processing of the resulting database. Though the target of research in formal semantics should be a complete formalization of semantic descriptions, Definiens will represent a significant advance in the field by offering access to the semantics of French that will go far beyond what is now available (which is simple processing of definitional texts as strings of characters).

2.3 A heterogeneous corpus of definitions

Identifying the exact list of lexical units described in the TLFi in order to tag its definitions is not a straightforward task. The dictionary contains 54,281 main entries (i.e. polysemic units), in which are embedded 18,095 entries that correspond to morphological derivatives (e.g. FRUITARISME under FRUIT). Main entries also include 59,168 defined phrases that are of two types: some correspond to idioms (e.g. FRUIT DEFENDU under FRUIT), others correspond to collocations (e.g *vol* ***domestique*** under VOL). By *collocations*, we mean semi-idiomatic expressions, also called *semi-phrasemes* in Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995). For the project, we have automatically extracted from the TLFi a total of 271,164 definitions that describe the meaning of these three types of lexical entities — lexemes, idioms and collocations.

3. Segmentation into definitional components

3.1 Aims of the segmentation process

The first stage of the project—presently under development—consists in identifying the main definitional components, namely the central component (CC) and the optional peripheral components (CP, for *composante périphérique*). We illustrate this stage of processing with the definition of BROUETTE ('wheelbarrow') below:

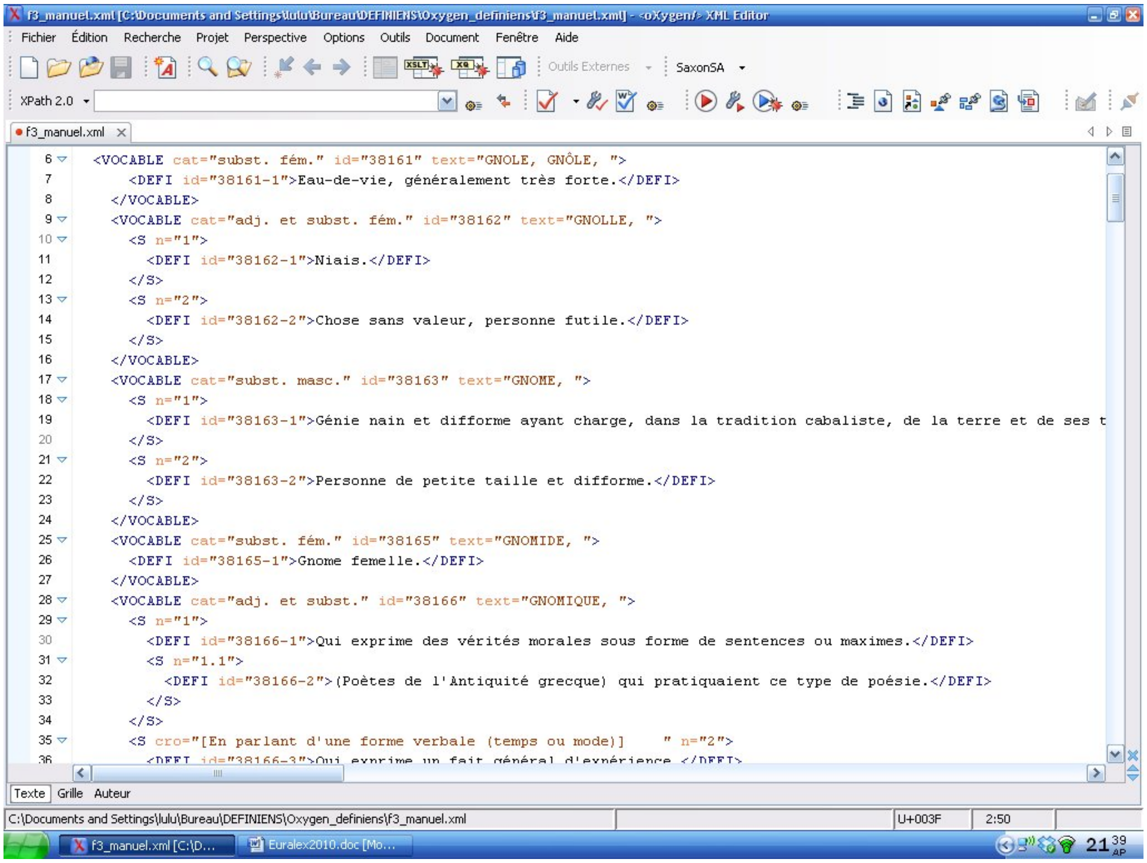
BROUETTE (sense B.1): *Véhicule à une roue et à deux brancards servant au transport des matériaux* 'Vehicle with one wheel and two handles, used to carry materials'

Identifying the central component of a definition is surely the most delicate task. In our example, one can hesitate between the component *véhicule* and the component *véhicule à une roue*. Since the central component has to be a classifying component and since the corpus of definitions does not include other lexical units described as a one-wheel vehicle, we choose the *véhicule* component as CC. We then easily identify the peripheral components, as illustrated below with the structured definition of BROUETTE: the PARAPH tag stands for 'paraphrase'.

> BROUETTE (sense B.1): <PARAPH><CC>Véhicule</CC> <CP>à une roue et à deux brancards</CP> <CP>servant au transport des matériaux</CP></PARAPH>

To date, about 5 percent of our corpus (that is, of 271,164 definitions) has been manually segmented. The manually segmented data are currently used to evaluate the automatic segmentation of the TLFi definitions and to describe the grammar of the TLFi metalanguage.

During their training period, the annotators work in pairs for the segmentation, and then alone. The agreement rate between annotators has not yet been evaluated. Annotators work on files (extracted from TLFi files and edited in the oXygen XML Editor) that contain only the relevant information for Definiens, that is, the name of the defined units and their definitions. The TLFi's structuration of the different meanings of the polysemic units (*vocable*) is preserved, as illustrated in figure 1 below.


f3_manuel.xml [C:\Documents and Settings\lulu\Bureau\DEFINIENS\Oxygen_definiens\f3_manuel.xml] - <oXygen/> XML Editor
Fichier Édition Recherche Projet Perspective Options Outils Document Fenêtre Aide
Outils Externes
SaxonSA
XPath 2.0
f3_manuel.xml
Texte Grille Auteur
C:\Documents and Settings\lulu\Bureau\DEFINIENS\Oxygen_definiens\f3_manuel.xml
U+003F
2:50
f3_manuel.xml [C:\D...
Euralex2010.doc [Mo...
21 39 AP

```xml
  <VOCABLE cat="subst. fém." id="38161" text="GNOLE, GNÔLE, ">
      <DEFI id="38161-1">Eau-de-vie, généralement très forte.</DEFI>
    </VOCABLE>
    <VOCABLE cat="adj. et subst. fém." id="38162" text="GNOLLE, ">
      <S n="1">
        <DEFI id="38162-1">Niais.</DEFI>
      </S>
      <S n="2">
        <DEFI id="38162-2">Chose sans valeur, personne futile.</DEFI>
      </S>
    </VOCABLE>
    <VOCABLE cat="subst. masc." id="38163" text="GNOME, ">
      <S n="1">
        <DEFI id="38163-1">Génie nain et difforme ayant charge, dans la tradition cabaliste, de la terre et de ses t
      </S>
      <S n="2">
        <DEFI id="38163-2">Personne de petite taille et difforme.</DEFI>
      </S>
    </VOCABLE>
    <VOCABLE cat="subst. fém." id="38165" text="GNOMIDE, ">
      <DEFI id="38165-1">Gnome femelle.</DEFI>
    </VOCABLE>
    <VOCABLE cat="adj. et subst." id="38166" text="GNOMIQUE, ">
      <S n="1">
        <DEFI id="38166-1">Qui exprime des vérités morales sous forme de sentences ou maximes.</DEFI>
        <S n="1.1">
          <DEFI id="38166-2">(Poètes de l'Antiquité grecque) qui pratiquaient ce type de poésie.</DEFI>
        </S>
      </S>
      <S cro="[En parlant d'une forme verbale (temps ou mode)]     " n="2">
        <DEFI id="38166-3">Qui exprime un fait général d'expérience.</DEFI>
```

Figure 1 : Edition in oXygen for the segmentation task

3.2 Automatic processing

3.2.1 The MACAON parsing tools

Segmenting and tagging the TLFi definitions is a time consuming and tedious task that can be partly automated using NLP techniques. We have devised a series of tools called MACAON that realizes the segmentation.

The segmentation process is decomposed into four consecutive sub-processes. The first one decomposes the definitional text into elementary elements called *tokens*, according to typographic rules. The second sub-process, called *lexer*, performs a search in a wide coverage inflected forms lexicon. During this step, tokens that have been identified by the tokenizer can be put together to form multi-word expressions. For example, the sequence *en dessous de* ('below') is recognized as a complex preposition. The output of the lexer is then passed on to the third sub-process, a part of speech tagger, whose output is passed on to the final sub-process, a chunker. The chunker groups POS-tagged words together to form complex units called *chunks* among which the chunks that are relevant for the segmentation work: the Central Component (CC), the Peripherial Components (CP) and the paraphrase (PARAPH).

The drawback of this sequential organization of the sub-processes is that most of them can, and generally do, make mistakes. An error made at one stage of the process usually provokes errors in the following stages. To address this problem, every module can take as input several solutions produced by the preceding module, and can produce several outputs. The chunker, for example, accepts several possible part-of-speech tag assignments to words and can produce several possible groupings of the tag sequences into chunks.

Ambiguous input and output are represented by way of weighted finite state automata and most modules process their input through standard operation on automata (Mohri, 1997). Each module also associates a score to every solution produced, and the scores of the different modules accumulate as long as the input sentence goes through the different modules. At the end of the process, every solution produced has a score which is a linear combination of the score given by each module that the solution passed through. The association of a score to each solution built by a module allows the researcher to limit, at any stage of the process, the ambiguity to the n highest score solutions. A scoring technique such as this has been shown by Nasr and Volanschi (2006) to improve the quality of a pipeline architecture. The highest score solution produced by the process is then reviewed by a human annotator.

3.2.2 First results

The automatic processing of the definitions is under development. The most delicate part of the development is the creation of the chunk grammar. Such a grammar is very ambiguous since there are generally many possible (syntactic) segmentations of a given definition. The ambiguity is further increased by the fact that several part of speech taggings of the definition are given as input to the chunker. We devised some simple heuristics in order to rank the possible segmentations and select the segmentation which achieves the highest score with respect to these heuristics. The current state of the chunk grammar allows us to parse 70.42% of a sample of the definitions. The quality of a segmentation H (for Hypothesis) of a definition D selected by the heuristics can be compared to the correct (manual) segmentation of D, which we call R (for Reference). H and R can be compared by means of recall and precision. The precision of H is the proportion of components (either central component or peripheral component) in H that are in R while the recall is the proportion of components of R that are found in H. The current state of the grammar achieves a precision of 0.74 and a recall of 0.58. In other words, 74% of the components automatically built are correct and 58% of the expected components have been built.

It is important to keep in mind that the aim of the automatic processing is to speed up the manual segmentation of the definition. In order to evaluate the efficacy of the automatic processing, a small experiment has been conducted. The annotators were asked to both annotate plain definitions and to correct automatically their annotated counterparts (and vice versa with another set of definitions). The time to perform both tasks was compared. The results are quite disappointing—at this stage, there is no noticeable gain in time: correcting the automatic segmentation is not faster than a manual segmentation from scratch. It is however most likely due to the quality of the automatic segmentation at this stage of the project; an automatic segmentation of better quality would surely yield some gains.

Several improvements are under study in order to increase the quality and coverage of the automatic segmentation and, hopefully, speed up the manual segmentation. The first one is the improvement of the chunk grammar and the heuristics used to rank the competing segmentations. The second one is more ambitious–it aims at segmenting all the definitions in parallel and not in an iterative way as is currently done. The idea is that a central component, for example, cannot be identified only on syntactic criteria, as is the case in the current system. Refer to our previous example, the definition of BROUETTE. In order to decide whether the central component of BROUETTE is *véhicule* or *véhicule à une roue*, we need to look up other definitions which have *véhicule à une roue* as a central component. The idea is to introduce this comparative search during the automatic segmentation stage: for each definition, all its potential central components are built based on syntactic criteria, but the choice of the best one is made after a comparison with the candidates of all the other definitions in the database. In our example, the candidate *véhicule à une roue* will not be selected since it does not appear as a potential central component of any other definition.

4. Semantic enhancement

4.1 Aims of semantic tagging

Semantic tagging first consists of assigning the central component a semantic label, i.e. a normalized expression that accounts for the semantic value of the central component. Every semantic label is defined by its appropriate definition in the TLFi, as illustrated below with the semantic label that will tag the central component of the definition of BROUETTE:

> véhicule (from VÉHICULE sense II.A): engin constitué d'un châssis muni de roues, à traction animale ou autopropulsé, servant au transport routier ou ferroviaire ‘machine made up of a chassis equipped with wheels, pulled by animals or selfpropelled, used for road or rail transport’

The definition of the semantic label guides us in the second step of the semantic tagging, namely the assignment of a definitional role to each peripheral component of a given definition. Indeed, the definition of VÉHICULE given above indicates that a vehicle has some “characteristic parts” (*constitué d'un châssis muni de roues*), is characterized by a “type of motion” (*à traction animale ou autopropulsé*) and, finally, has a “function” (*servant au transport routier ou ferroviaire*). This tells us that these three basic roles will likely be found in the definition of lexical units labeled with *véhicule* (or a more specific label). An examination of the definitions of the set of lexical units labeled *véhicule* reveals indeed other roles like “speed”, “appearance”, etc. The definition of BROUETTE will hence be semantically tagged as follows:

> BROUETTE (sense B.1): <PARAPH><CC=**véhicule**>Véhicule</CC> <CP=**parties charactéristiques**>à une roue et à deux brancards</CP> <CP=**fonction**>servant au transport des matériaux</CP></PARAPH>

4.2 A closer look into the hierarchy of semantic labels

The hierarchy of semantic labels, developed during the construction of the DiCo database (Polguère, 2003), currently contains 790 labels that have been defined with the relevant definition from the TLFi, as illustrated by *véhicule* in the previous section. As these labels have been created for the description of a limited but representative set of lexical units, we do not as yet know to what extent the labels will cover the TLFi word list.

An important lexicographic work has to be done to enhance the label hierarchy with roles. As previously mentioned, each label is associated a list of possible roles for predicted peripheral components, as illustrated below with *véhicule*:

véhicule: {parties caractéristiques, fonction, mode de fonctionnement, vitesse, apparence}

This kind of information will be very useful for the continuation of our project, which will consists of creating a new lexicographic database made of fully formalized definitions. Indeed, the set of roles associated with each label allows for a semantic control of the definitions. For example, a semantically well-formed definition of a lexical unit labelled *véhicule* has to contain peripheral components whose roles are associated to this label in the hierarchy.

5. Conclusion

In this paper, we have presented an ongoing lexicographic project which aims at providing an explicit and formalized structuring for the definitions of the French electronic dictionary *Le Trésor de la Langue Française informatisé* (TLFi). The project is ambitious and will take significant time to be completed. In the meantime, the following intermediary results can be achieved. First of all, the basic segmentation of the definitions into central and peripheral components will allow for a better exploitation of the dictionary. For instance, one will be able to request the set of lexical units that denote a vehicle with two wheels, or the set of lexical units that denote an "intense" feeling. This type of request is very useful for a language-related task which requires well specified semantic clusters of lexical units (for example, information extraction, etc.). The work in progress will also lead to a precise characterization of the TLFi metalangage and will thus allow for the description of a definition format in order to develop an NLP lexicon derived from the TLFi. This could be done using the *Lexical Markup Framework* (LMF), for example (Francopoulo, 2005:19-21). Finally, the semantic label hierarchy, which will probably be completed before the end of the project, will offer a valuable resource to lexical semantics researchers.

Acknowledgements

This research is supported by a grant from the Social Sciences and Humanities Research Council (SSHRC) of Canada and by an internal grant from the ATILF CNRS (Nancy) research lab. We wish to thank our invaluable Canadian and French team of annotators: Évelyne Arbour-Barbeau, Jean-Luc Benoit, Claudia Fecteau, Anaïs Ferte, Anne-Laure Jousse, Caroline Poudrier, Mélissa Samson-Fauteux and Olivier Taïs. Finally, we are deeply grateful to Helen Lim for proofreading the final version of this paper.

References

Altman J., Polguère A. (2003) La BDéf : base de définitions dérivée du *Dictionnaire explicatif et combinatoire*. *Proceedings of the First international conference on the Meaning-Text Theory (MTT'2003)*, Paris, 43-54.

Aristotle (1939, translation by E. S. Forster) *Topica*. Cambridge, Mass.: Harvard University Press.

Barque L. (2008) *Description et formalisation de la polysémie régulière du français*. Ph.D. dissertation, Université Paris 7, Paris.

Dendien J., Pierrel J.-M. (2003) Le Trésor de la Langue Française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues (T.a.l.)*, 44(2):11-37.

Francopoulo, G. (2005) Extended examples of lexicons using LMF (auxiliary working paper for LMF). Rapport technique, INRIA-Loria.

Mel'čuk I. *et al.* (1984, 1988, 1992, 1999) *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques I–IV*. Montréal: Les Presses Universitaires de Montréal.

Mel'čuk I., Clas A., Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Mohri M. (1997) *Finite-State Transducers in Language and Speech Processing*, Computational Linguistics 23(2).

Nasr A., Volanschi A. (2006) *Integrating a Part of Speech Tagger and a Chunker Implemented as Weighted Finite-State Machines*, Finite-State Methods and Natural Language Processing 167-178.

Martin R. (1979) La polysémie verbale, esquisse d'une typologie formelle. *Travaux de linguistique et de littérature*, 17:261 276.

Polguère A. (2003) Étiquetage sémantique des lexies dans la base de données DiCo. *Traitement Automatique des Langues* (*T.a.l.*), 44(2):39-68.

Polguère A. (2008) *Lexicologie et sémantique lexicale. Notions fondamentales*. Montréal: Les Presses de l'Université de Montréal.

Pustejovsky J. (1995) *The Generative Lexicon*. Cambridge, Mass.: MIT Press.

Rundell M. (2008) More Than One Way to Skin a Cat: Why Full-Sentence Definitions Have Not Been Universally Adopted. In T. Fontenelle (ed.): *Practical Lexicography. A Reader*. Oxford, UK: Oxford University Press, 197-209.

Sinclair J. M. (ed.) (1990) *Collins Cobuild Student's Dictionary*. London: Collins.
TLFi (2004) *Trésor de la Langue Française informatisé*. Paris: CNRS Éditions. [Online access: http://atilf.atilf.fr/tlf.htm]

Wierzbicka A. (1987) *English Speech Act Verbs. A Semantic Dictionary*. Sydney *et al.*: Academic Press.