Construction of a French Lexical Network: Methodological Issues

Veronika Lux-Pogodalla[†], **Alain Polguère**[*†]

[†] ATILF CNRS [*] Nancy-Université
44, avenue de la Libération, B.P. 30687
54063 Nancy Cedex
France
{veronika.lux, alain.polguere}@atilf.fr

Abstract

We present a new lexicographic enterprise that aims at producing a *French Lexical Network* or *FLN*. We begin by introducing the project as such and then proceed with a characterization of the FLN: the FLN as a generic lexical model, its network structure and the different types of lexical entities it models. Finally, we focus on two aspects of our lexicographic methodology: the incremental identification of the FLN's wordlist and our editing tool.

1. The French Lexical Network project

We present a lexicographic project that has just been officially launched (early 2011) and whose aim is to build a new type of lexical resource called *French Lexical Network*, hereafter *FLN*.[1] Though the construction of the FLN is conceived as a long-term enterprise, we focus here on the first three-year phase, i.e. the tasks that have actually been planned and funded in the context of a more global R&D project called *RELIEF*. For lack of space and in order to concentrate on the very specific topic of lexical resources' design and construction, we will ignore the application/valorisation aspects of RELIEF and exclusively deal with the FLN itself.

We are fully aware of the fact that, by presenting a lexical resource that is only emerging from the drawing board, we have no tangible "results" to offer as yet. However, we believe that the FLN project is sufficiently specified, both in terms of design of the lexical resource and of lexicographic methodology, to be of interest for the research community—not to mention the importance there is for the FLN team to benefit from early feedback from this community. Additionally, it will appear clearly in what follows that the FLN is not a project that started from scratch, but a project that directly builds on previous research and lexicographic work performed over the last two decades.

The structure of the remainder of the paper is as follows. The main characteristics of the FLN are presented in section 2.: the FLN as a generic lexical model, its network structure and the different types of lexical entities it describes. In section 3., we focus on two aspects of our lexicographic methodology: the incremental identification of the FLN's wordlist and our editing tool.

Before we begin, let's mention that a lexicographic team of around 15 persons is being put together for the initial three-year phase of the FLN project. Lexicographic strategies embedded in our theoretical and methological framework of reference—the Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995; Mel'čuk, 2006)—will serve to extract linguistic information from corpora. However, we will also make extensive use of the *Trésor de la Langue Française informatisé* (Dendien and Pierrel, 2003), hereafter *TLFi*,[2] as a mother lexical database from which we will extract lexicographic information to be reinterpreted and exploited within the FLN.

2. Main characteristics of the FLN

This section offers a three-step characterization of the global structure of the FLN: the FLN as a generic lexical model (2.1.), the lexical network structure of the FLN (2.2.) and the various types of lexical entities this network connects (2.3.).

2.1. Generic lexical model

In a nutshell, the FLN is designed to belong to the *-Net* family of lexical resources, such as WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 2003; Ruppenhofer et al., 2010). In addition to its network structure, that will be examined in section 2.2., it shares two main characteristics with WordNet and FrameNet:

1. it is **not** a dictionary, i.e. it is not a "textual," linear model of the lexicon;
2. it is nevertheless built in a lexicographic way, i.e. manually by a lexicographic team.

Like WordNet and FrameNet, the FLN has been from the onset conceived of as a generic, general purpose lexical database. However, it is possible to derive a wider range of lexical resources from it: lexicons for NLP systems, full-fledged dictionary entries (similar to those of standard dictionaries) and on-line lexical resources for language teaching/learning. For this reason, it is not focusing on a limited set of specific properties of lexical units (such as semantic hierarchical organization of synsets for WordNet or frames controlling the semantics-syntax interface for FrameNet), but adopts a global view of all lexical properties, very much

[1] The French name of the targeted lexical resource is *Réseau Lexical du Français* or *RLF*

[2] *TLF* stands for the original "paper" dictionary and *TLFi* for its electronic on-line version. The TLFi's URL is: `http://atilf.atilf.fr/tlf.htm`.

like dictionaries would do: lexicographic definition, grammatical features, syntactic combinatorics (roughly, subcategorization frames), lexical combinatorics and derivational links. In that respect, the FLN is equivalent to a virtual dictionary (Selva et al., 2003) or, rather, to virtual dictionaries of various macro- and microstructures that can automatically be generated from it.

2.2. Lexical network structure

The FLN's architecture is similar to a lexical system, as presented in Polguère (2009): a huge network of lexical units connected by a broad range of lexical links encoding semantic or combinatorial lexical relations. The bulk of the network structuring is carried out by means of the system of **standard lexical functions** (Mel'čuk, 1996), that allows for a rigorous encoding of lexical paradigmatic links (synonymy, antonymy, conversivity, actant names, etc.) as well as syntagmatic links (collocations controlled by lexical units—their typical intensifiers, support verbs, etc.). Lexical functions have previously been used in the design of other lexical databases (Fontenelle, 1997; Selva et al., 2003); the FLN is drawing mainly from previous work done on the **DiCo** lexical database (Steinlin et al., 2005) in making use of a double encoding of lexical links:

1. formulas based on the formal language of lexical function relations (Kahane and Polguère, 2001);

2. "popularization" of these formulas in the form of paraphrases (in controlled natural language) of the corresponding paradigmatic or syntagmatic link.

For instance, following this approach, the paradigmatic link holding between KILL$_V$ [*These mushrooms can kill you!*] and LETHAL is to be encoded as follows in the lexicographic article for this sense of KILL$_V$ (popularization comes first, followed by the lexical function formula):[3]

[X] that can $\sim$
Able$_1$ *lethal*

The formal encoding allows for various computations on the lexical graph and the popularization allows for the generation of general public lexicographic descriptions (dictionary articles) from the lexical database.[4]

Beside lexical-functional links, the FLN graph will also encode embedding of semantemes (lexical senses) through its formal definitions—see section 3.2.2. below.

The main aim of the FLN network structuring is to build a model of French lexical knowledge that is truly generic and independent of any specific textual (dictionary-like) or hierarchical (ontology-like) organization. It can also be expected that the chosen model, because of its non-textual nature, will be closer to what is generally believed to be the network-like structure of the mental lexicon (Aitchison, 2003). The main originality of the FLN in terms of structuring, when compared to databases of the *-Net* family, is that it proposes a multi-dimensional graph structure for all standard paradigmatic and syntagmatic links; it does not organize lexical information "through the eyes" of just a few selected links, such as hyperonymy or synonymy. To the best our knowledge, such structure has yet to be implemented, at least for the French language.

2.3. Lexical entities that are nodes of the FLN graph

The FLN will be stored as an SQL database, which will implement its network structure as a set of connections between lexical entities of different types. Central to the lexicographic description are **lexical units** proper, which are of two kinds:

1. **Lexemes** are monolexemic lexical units such as Fr. COUP **I.1** [*Il a reçu un **coup** sur la tête en tombant.*][5] or COUP **I.2** [*Le voleur lui a donné un **coup** sur la tête.*].[6] They correspond to so-called *word senses*.

2. **Idioms** are syntagmatic lexical units such as Fr. ⌜COUP DE SOLEIL⌝ 'sunburn' (lit. 'knock of sun').

Only lexemes and idioms are considered in the FLN as full-fledged lexical units, and they are the actual units of lexicographic description. **Vocables**—polysemic words—are modelled as sets of lexical units connected in the graph by a relation of copolysemy.

The FLN will put strong emphasis on phraseology, i.e. on the set phrases of the language, known as **phrasemes**. Following Mel'čuk (1995), three main types of phrasemes are being considered: (full) idioms, linguistic clichés and collocations.

Because they are lexical units, as much as lexemes are, **idioms** will be described by "normal" lexicographic articles, and not embedded in the article of one of the lexemes they formally contain. For instance, ⌜COUP DE SOLEIL⌝ is not to be described as embedded lexical entity in the article for COUP **I.1**, as it is presently the case in standard language French dictionaries such as *Petit Robert* (Rey-Debove and Rey, 2010).

Linguistic clichés, such as Fr. *Après vous !* 'Go ahead!' (lit. 'After you') are the second type of phrasemes that will be accounted for by lexicographic articles. However, because they are not actual lexical units, clichés will not be considered as "entries" in the database and will receive a somewhat simplified description: no actual lexicographic definition (which will be replaced with the specification of the communicational goals of the speaker) and no indication of combinatorial properties.

As for **collocations**—compositional though phraseological expressions (Hausmann, 1979; Benson et al., 1997)—, they will be accounted for in the article for their base by means of syntagmatic lexical functions, following the approach taken in the DiCo (already mentioned in section 2.2.) and other related lexicographic models.

[3] *X* stands here for the first actant of the keyword (KILL$_V$ = 'X kills Y') and $\sim$ for the keyword itself.

[4] For a general public dictionary (manually) generated from the DiCo database, see the *Lexique actif du français* (Mel'čuk and Polguère, 2007).

[5] *He got a **knock** on his head when he fell.*

[6] *The burglar stroke him a **blow** on his head.*

It can be noted that the lexicological principles adopted for the FLN are very much the same as those of the DiCo project, except for two major differences:

1. Each lexical unit is to be semantically described by a complete and formalized lexicographic definition—whereas the DiCo only provides a description of the actancial structure of the unit together with a semantic label (Polguère, 2003; Polguère, To appear).

2. The data structure of the FLN is a true lexical system, i.e. a network of semantic and combinatorial connections between lexical units. The DiCo's lexical links are in reality connecting lexical units to string of characters (lexical forms), pretty much like any standard dictionary.[7]

By reifying the target of lexical links, the FLN will play in the same "formal" league as WordNet or FrameNet—though its initial vocabulary coverage will of course be very small in comparison (see section 3.1. below, on the FLN's coverage).

3. Lexicographic methodology

This section deals with two methodological aspects of the project that we consider crucial and to which particular attention has been paid: the incremental identification of the FLN's "wordlist" (3.1.) and the writing of FLN articles (3.2.3.).

3.1. The FLN's lexical coverage

3.1.1. Incremental identification of the "wordlist"

In the long run, the FLN should cover the bulk of basic contemporary French. This is a gigantic task, that can only be handled through a series of carefully planned successive efforts. As mentioned earlier, this paper deals exclusively with the initial three-year phase. At the end of this first phase, the FLN should possess a "wordlist"—though the term *wordlist* may not be fully relevant in the specific case of a lexical network—of at least 10,000 vocables.[8] How are these vocables selected among the 70 to 80,000 vocables described in a standard commercial dictionary such as *Nouveau Petit Robert*, idioms included?

The FLN is not designed as a dictionary and, therefore, the process of selecting and building the wordlist can be very different from the selection process implemented by lexicographers of "traditional" dictionaries, such as the TLF, our dictionary of reference (see end of section 1. above). Because of publishing constraints—need for regular releasing of fully completed volumes—, the TLF lexicographers had to first define a whole wordlist, proceeding afterwards through it in strict alphabetical order: vocables starting with the letter *A* being described first, those with letter *B* second, etc. Contrary to this, our progression will not be alphabetical. It will proceed through series of important lexical fields of the language: vocables whose basic lexical unit belong to the semantic field of feelings, of relationships, of animals, of tools, etc. This allows us to start with an initial priming wordlist—Fr. *nomenclature d'amorçage*—that will constantly grow during the project, following a logic that will be detailed shortly.

3.1.2. The priming wordlist

How do we determine the priming wordlist that will be the "seed" from which the whole FLN wordlist will grow in the years to come? In the beginning, priority is given to the most basic, common French vocables. To identify them, we made use of four types of sources:

1. well-known lists of "basic French" developed mainly for applications in language teaching; essentially: the 3,500 vocables of the *Français fondamental* (Gougenheim et al., 1967) and the 3,787 vocables of the *Échelle Dubois-Buyse* (Ters et al., 1988);

2. the "Éduscol" vocabulary list of the 1,462 most frequent lemmas found in the 19th and 20th century French literature;[9]

3. the 6,500 vocables wordlist of the *Robert Benjamin* (Collectif Robert, 2009), a very high quality and seasoned pedagogical French dictionary used in primary schools;

4. a vocabulary wordlist of 4,548 lemmas compiled at the Université de Montréal for the Quebec ministry of education (Ministère de l'Éducation du Loisir et du Sport, MELS) using a meticulous and well-specified methodology (Lefrançois et al., 2011).

Through a cross-checking process,[10] we have identified a priming wordlist of 3,739 vocables, which we believe will induce the description of the basic, minimal set of vocables any speaker of the language, any NLP system, etc., should master.

The number of 3,739 may seem arbitrary, and to some extent it is. This, however, is inconsequential for three reasons. First, it can be noted that most studies on vocabulary thresholds for basic language proficiency conclude to vo-

[7] Of course, a lexical link in the DiCo can specify the actual lexical sense that is the target of the link (`coup#I.1` instead of just `coup`). This, however, is only transparent for the human user of the database and no actual connection is implemented at the level of the data structure.

[8] In comparison, the DiCo—which covers a "sample" rather than a "core" French vocabulary—has a wordlist of 395 finalized (status 0) and 145 prefinalized (status 1) vocables, for a total of 1,127 word senses. The DiCo is accessible on-line in two forms: 1) the *DiCouèbe* interface to DiCo's SQL tables (`http://olst.ling.umontreal.ca/dicouebe`) and 2) the *DiCoPop* dictionary pages automatically generated from the SQL tables (`http://olst.ling.umontreal.ca/dicopop`).

[9] This list, compiled at the Institut National de la Langue Française (INaLF), is available from the Éduscol French government website: `http://eduscol.education.fr/`.

[10] For instance, the *Robert Benjamin*'s wordlist contains many vocables that are mainly relevant in the context of primary school education and by no means belong to the minimal core of French vocabulary—QUADRILATÈRE 'quadrilateral$_N$', SORCIER/SORCRCIÈRE 'sorcerer'/'sorceress,' etc. Such vocables are not to be included in the priming wordlist.

cabulary sizes that range from 3,000 "word families"[11] for basic use to 9 to 10,000 for advanced proficiency (Hirsh and Nation, 1992; Nation, 2006). Our 3,739 vocables priming wordlist is therefore in the lower bracket, but still in the realm of what can be considered as a reasonable, basic vocabulary. Second, what matters most is that the vocables we have selected do all belong to basic French and none are peripheral elements of the French vocabulary. Third, it is irrelevant whether one, or two, or 36 vocables have been omitted whose inclusion in the priming wordlist vocabulary would be justified. If a vocable is "missing" for whatever reason, and if it truly belongs to basic French, the induction process that we are now about to describe will catch up with it and have it included in the induced wordlist—Fr. *nomenclature induite*.

3.1.3. The induced wordlist

There are three different ways a vocable that is not present in the priming wordlist can be induced from it: 1) its basic lexical unit is a "close" semantic derivative (nominalization, verbalization, etc.) of the basic lexical unit of a priming vocable, 2) it is a very common idiom formally made up of lexemes of the priming wordlist or 3) its various senses are the target of a significant number of lexical links originating from the lexicographic description of units of the priming wordlist.

1) Induced close semantic derivatives A lexical unit L_2 is a semantic derivative of a lexical unit L_1 if it is the target of a paradigmatic lexical-functional link originating from L_1. The semantic derivation relation between these two units may or may not be marked morphologically.

We use the eleven following paradigmatic lexical-functional links to identify what we term the **close semantic derivatives** of a given lexical unit L.

1. **Syn**: exact synonyms of L, e.g. MOVIE → FILM_N;
2. **Anti**: exact antonyms of L, e.g. LEGAL → ILLEGAL;
3. **of opposite sex Syn$_\cap$**: quasi-synonym (more specifically, intersecting synonym) of L that denotes the same individual/animal as L but of the opposite sex, e.g. ACTOR → ACTRESS, DOG → BITCH;
4. **$\mathbf{V_0}$**: verbal conversion of L, e.g. KNOCK_N → KNOCK_V;
5. **$\mathbf{S_0}$**: nominal conversion of L, e.g. KNOCK_V → KNOCK_N;
6. **$\mathbf{Adj_0}$**: adjectival conversion of L, e.g. COAST_N → COASTAL;
7. **$\mathbf{Adv_0}$**: adverbial conversion of L, e.g. SLOW_Adj → SLOWLY;
8. **$\mathbf{S_i}$**: nouns meaning 'i$^{\text{th}}$ actant of L', e.g. DRIVE_V → DRIVER [= **$\mathbf{S_2}$**];
9. **$\mathbf{A_i}$**: adjectives meaning 'that is the i$^{\text{th}}$ actant of L', e.g. HUNGER → HUNGRY [= **$\mathbf{A_1}$**];
10. **$\mathbf{Able_i}$**: adjectives meaning 'that has the ability to be the i$^{\text{th}}$ actant of L', e.g. LOVE_V → LOVABLE [= **$\mathbf{Able_2}$**].
11. strict **Mult**: collective nouns that do include in their definition the meaning of L, e.g. LEAF → FOLIAGE—but SCHOOL [*of fish, shrimps...*] is not induced directly from FISH, as it is too vague.

Notice that the eleven above-mentioned lexical functions are used here in their "narrow sense," described in the glosses that accompany them. For instance, strictly speaking, VICTIM [*of a murder*] is a valid **$\mathbf{S_2}$** for MURDER_N [*by X of Y*], but it should not be considered as being a close semantic derivative because its meaning is much vaguer than 'Y of a murder' (**murderee*).

It is good practice in Explanatory Combinatorial Lexicography to describe a vocable V together with all vocables whose basic lexical unit (basic sense) is a close semantic derivative of the basic lexical unit of V. For instance, MURDER_N should necessarily be lexicographically described together with MURDER_V, MURDEROUS, MURDERER and MURDERESS. In order to adhere to Explanatory Combinatorial methodology, we consider as induced vocables all vocables whose basic lexical unit is a close semantic derivative of the basic lexical unit of a priming vocable. For instance, though PRÉVISION 'prediction' is not in our priming wordlist, it is included into the induced wordlist as it is a close semantic derivative of the priming vocable PRÉVOIR 'predict.' Notice however that, at this stage, only close semantic derivatives that are commonly used and do not belong to specialized vocabularies will be induced. For instance, HASE 'femal hare' is a close semantic derivative of LIÈVRE '(male) hare,' which belongs to the priming wordlist, but it will not be directly induced from it because of its almost technical nature.

2) Induced idioms The priming wordlist is made up of lexemic vocables. Any common idiom that is formally made up of lexemes that belong to the priming wordlist will be systematically included into the induced wordlist. For instance, as COUP, DE and SOLEIL (see section 2.3. above) all belong to the priming wordlist, ⌜COUP DE SOLEIL⌝ 'sunburn' is identified as induced vocable and added to the lexicographic team's in-tray.

3) High degree nodes of the graph In the process of describing vocables of the priming wordlist, lexicographers will be lead to introduce "on the fly" many new nodes in the FLN graph. They correspond to lexical units that are the target of links originating from priming lexical units. Two main types of links have to be considered.

Firstly, any lexical unit used in a lexicographic definition for a priming lexical unit is necessarily the target of a lexical link (of semantic inclusion). If this target is itself a priming lexical unit, nothing needs to be done. If it is not, a minimal entry for it is generated on the fly in order to make the link hold[12] (on FLNs' definitions, see section 3.2.2. below). For instance, if ASTRE—a very basic but not so com-

[11] In P. Nation's terminology, a word family is a word morphological base form plus all its associated inflectional variants and regular morphological derivations.

[12] Of course, it can also be the case that this unit, though not priming, is already present in the lexical graph as a result of an earlier on-the-fly generation.

mon term roughly equivalent to 'celestial body'—is used as a generic component in the definition of the lexeme SOLEIL 'sun,' then ASTRE will be included in the FLN graph, with minimal information (mainly, its part of speech and some illustrative linguistic examples).
Secondly, any lexical unit that is the target of a lexical-functional link originating from the description of a priming lexical unit also has to be inserted on the fly in the FLN graph. For instance, the adjectival unit RETENTISSANT 'resounding' will be inserted in the graph though it is not among the 3,739 units of the priming wordlist because COUP **I.1** 'knock' is a priming lexical unit and RETENTISSANT is one possible **Magn** (= intensifier) for it.[13]
As a result of the strategy of on-the-fly creation of entries for targeted lexical units, the FLN graph will gradually incorporate a large number of roughly sketched nodes that did not belong to the priming wordlist. A statistical analysis of the graph will regularly be performed in order to identify a list of top non-priming nodes of the graph that possess a high degree of connectivity. These nodes define the next batch of vocables to be inserted in the induced wordlist.
Figure 1 visualizes the wordlist expansion via insertion of idioms, close semantic derivatives and targeted units.

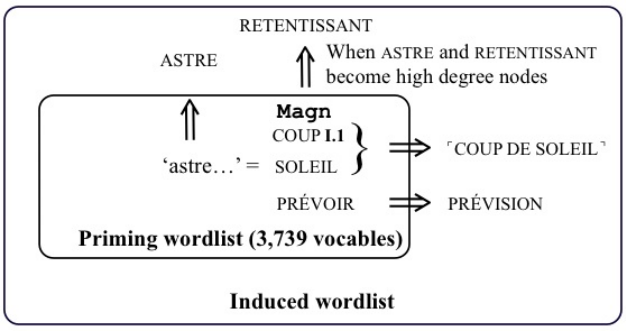

Figure 1: Self induced expansion of the wordlist

As we see, starting from the initial priming wordlist, the FLN will induce its own expansion according to a very simple logic: lexical units that are often referred to by units of the priming wordlist are "important" units, on which lexicographic work should focus. This strategy can be applied indefinitely as a guide to the expansion of the FLN.

3.2. Writing of FLN's articles

3.2.1. In lexicography, size matters

Work previously done on the DiCo database and on other extremely rich and formalized lexicographic models based on Explanatory Combinatorial Lexicology (Mel'čuk et al., 1984 1988 1992 1999) has shown that even skilled lexicographers fail to ensure the coherence of such lexical models when they grow to more than a thousand entries or so. If one wants to use this kind of approach to embark on a major lexicographic project, a rich tailor-made editing environment is required.
FLN articles have to comply to a well-specified structure that could be encoded, in theory, as an XML schema, and enforced through the use of an XML editor. However, there are two aspects of the FLN project that make it impossible to rely on such basic lexicographic tools.
Firstly, building the FLN is a true, large-scale lexicographic enterprise involving the coordinated work of an organized lexicographic team. There is therefore a need to possess an editor that, on top of ensuring the control of the formal validity of the description, will implement a lexicographic production line, with its various tasks (drafting, development, completion with corpus data, revision cycles, etc.) and their logical organization—a workflow management tool system.
Secondly, what really makes the editing of an FLN article complex is the fact that the information it contains has to be stored not as text, but as a database of connected entities forming a lexical graph. We believe that only this type of data structure will ultimately allow us to perform efficient consistency checks and other logical operations on our model of the lexicon. We are particularly interested in the possibility of using the graph structure of the FLN and formal properties of lexical-functional links to implement semi-automatic drafting of vocables based on potential analogies with already existing descriptions—on this, see Jousse (2010, p. 236–257).
Off-the-shelf professional dictionary production softwares such as TLex[14] (de Schryver and de Pauw, 2007) do exist and are used to build major commercial dictionaries. In our case, we chose to work in close collaboration with MVS Publishing Solutions,[15] our partner in the RELIEF project (see section 1.), to tune their Dixit editor for our specific needs. This editor is a component of a software suite mainly used for the publication of daily newspapers. It controls the writing process of newspaper articles (structuring of the article, handling of its editorial cycle and SQL storage of textual as well as non-textual information), data management and automatic generation of printed articles based on predefined layout rules. Thus, it already contains all functionalities one needs in order to perform the writing and, even, publication/dissemination of lexicographic articles.
In the remainder of this section, we will first describe the FLN microstructure the editor has to handle (3.2.2.), then explain the main features of the editor (3.2.3.).

3.2.2. Structure of a lexicographic article

The structure of an FLN article is very similar to that of a DiCo record (Lareau, 2002; Jousse and Polguère, 2005), and an SQL export of the DiCo data is actually being used for tuning the FLN lexicographic editor. As can be seen in Figure 2 below, with the article for ADMIRER **I** 'to admire [someone for something],' an FLN article is divided into six main sections, of which only the second one–Definition–is absent from DiCo records and will therefore be presented in some detail here.

1. **Grammatical features** This section lists features encoding combinatorial properties of the keyword (register, part of speech, inflectional restrictions, etc.).

[13] More precisely, it corresponds to the semi-standard lexical function **in respect to noise Magn**, or $\mathbf{Magn}_{\mathbf{noise}}$.

[14] http://tshwanedje.com/
[15] http://www.mvs.fr/

2. **Definition** In the FLN, each full lexical unit is to be semantically described by means of a paraphrastic definition (which was not the case in the DiCo). Each definition is made up of two components:

   a. The *definiendum* is a description of the actancial structure of the keyword.

   b. The *definiens* (definition proper) is the analytical paraphrase of the keyword's meaning. Prototypically, a definiens is mainly made up of a central component (CC) and one or more peripherical components (PC). Lexicographers annotate the text of the definiens so as to make its internal structure explicite. For example, the definiens in Figure 2 below is encoded in the background as follows:[16]

```
<DEFINIENS label="apprécier">
     <CC>L'individu X apprécie Y pour Z</CC>
     <PC role="intensity"> beaucoup</PC>
     <PC role="cause"> du fait des qualités
         exceptionnelles de Z</PC>
</DEFINIENS>
```

   As indicated in 3.1.3., each lexical item occurring in the definition is connected by a semantic inclusion link to a specific lexical unit—whether priming, induced or pending description—, whose own definition, if it exists, will be subjected to the same formal treatment. Of course, such strategy will make the process of writing a lexicographic definition very slow and, in some respects, tedious. It should be noted, however, that it has the positive effect of forcing lexicographers to proceed very selectively and with economy in writing lexicographic definitions, thus ensuring the production of definitions of greater clarity[17]—see, for instance, the systematic use of a basic defining vocabulary in the definitions of the Longman dictionary (Summers, 2005).

3. **Government pattern** This section describes how the keyword's semantic actants can be expressed as its syntactic dependents. A database of French government patterns will be included in the FLN data structure and valency tables (roughly, subcategorization frames) appearing in a lexicographic article will ultimately be directly imported from this base rather than manually typed by lexicographers.

4. **Lexical functions** This section is the core of the lexical description, as explained in 2.2. Lexical links implemented here will be the main structuring elements of the FLN lexical graph. For lack of space, we cannot enter into the details of the encoding of paradigmatic and syntagmatic links by means of lexical functions. This topic is largely dealt with in the literature on Explanatory Combinatorial Lexicology cited in this paper.

5. **Examples** This section of FLN articles will be much more structured than what can be seen in Figure 2, where only examples imported from the DiCo appear. In an actual FLN article, there will be several types of lexicographic examples, mainly: citations from texts of various genres with exact references—extracted from Frantext[18] and other ATILF in-house corpora—and hand-crafted adaptations of corpus/Internet data.

6. **Phraseology** This last section lists idioms or linguistic clichés that formally contain the keyword. Each enumerated phraseme is linked to the corresponding FLN article.

3.2.3. Designing a lexicographic editor

Recall that the unit of lexicographic description in the FLN is the lexical unit: lexeme ("word" taken in one specific sense) or idiom. Though other lexical entities—such as linguistic clichés (cf. 2.3. above)—may be described by means of lexicographic articles, the editor is essentially providing an interface for lexicographers to describe properties of lexical units.

The lexicographic editor for FLN is currently being prototyped by MVS Publishing Solutions using their Dixit general purpose editor. Figure 2 below is a sample screendump of the editor's interface in its present, very preliminary state. It shows the ADMIRER I entry, based on DiCo data to which a full-fledged lexicographic definition has been added. The purpose of this figure is mainly for the reader to visualize better the type of lexicographic data we are dealing with.

The editing interface helps lexicographers produce descriptions that comply to the microstructure presented above. Practically, their task is closer to filling-in a very complex and structured form than to performing free writing, which is precisely what is required for lexicographic tasks. Moreover, in each section, the editor provides assistance to control compliance to particular constraints on content, ultimately ensuring that the entry is built as a valid subgraph of the global FLN. Depending on the constraints, the level and type of assistance will vary in each section, for instance:

- Normalized content can be directly selected from menus. Text items selected from menus are non-editable text in the article. (They can only be modified through menu selection.)

- Normalized content can also be selected via a form providing filtering features. This is for instance the case with the *Lexical functions* section: there are hundreds of potential lexical function formulas, too many for a single menu. Lexicographers can either indicate some features of the lexical function they are looking

[16] For more information on this approach to formally structuring lexicographic definitions, see Barque et al. (2010).

[17] The rather wordy definition for ADMIRER I in the TLF is *Considérer quelqu'un ou quelque chose avec un sentiment d'étonnement mêlé de plaisir exalté et d'approbation, le plus souvent motivé par la supériorité qu'on lui reconnaît dans divers domaines de la vie intellectuelle, esthétique, morale, etc.* 'To consider someone or something with a feeling of mixed exalted pleasure and approbation, usually motivated by the superiority one acknowledges to him/it in various aspects of life—intellectual, esthetic, moral, etc.'

[18] http://atilf.atilf.fr/frantext.htm

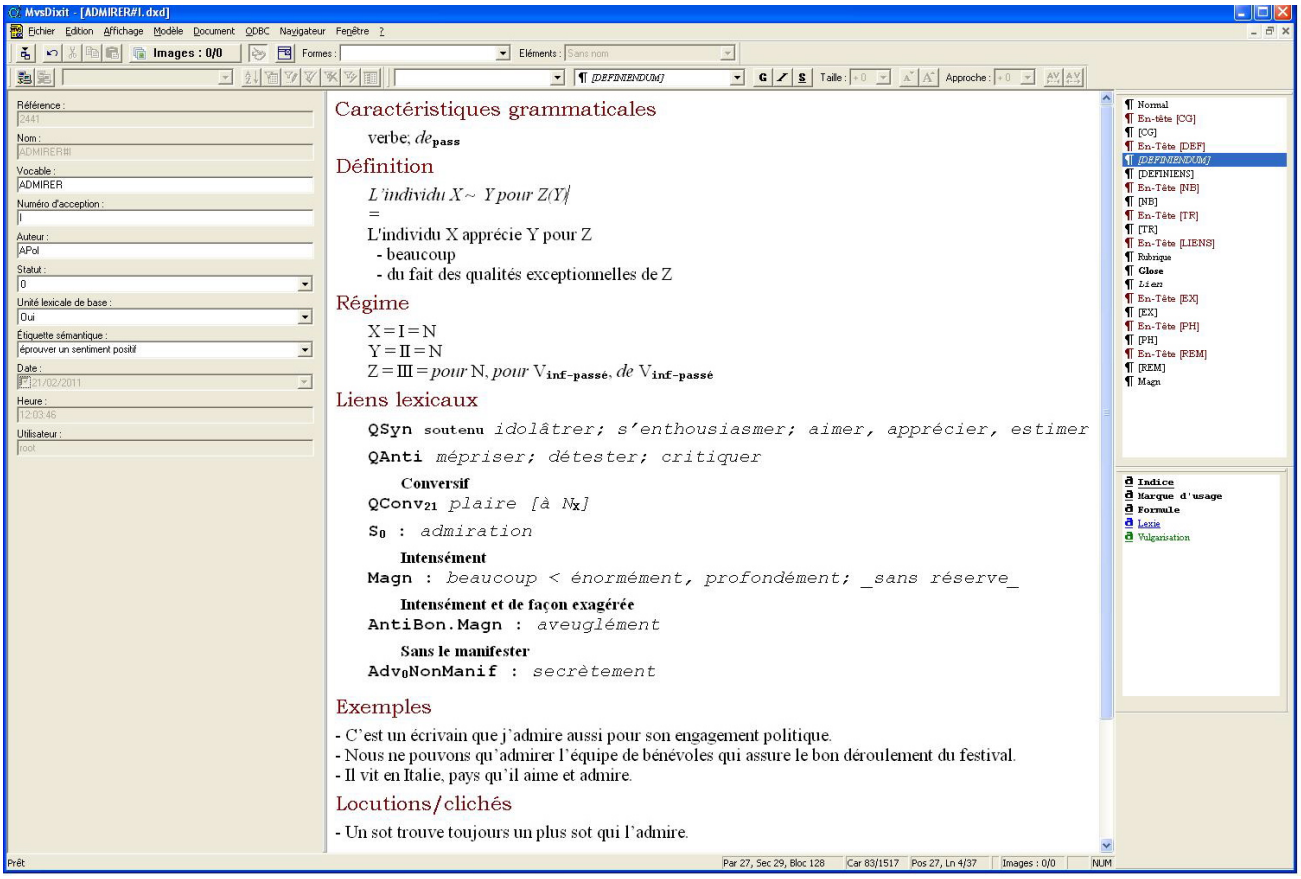

Figure 2: DiCo's data for ADMIRER **I** 'to admire [someone for something]' processed with the FLN editor

for (part of speech of the lexical target, etc.) in a form and get a filtered list of lexical functions in a menu, or they can start typing in the name of the function and get a list of suggestions (through a completion function). Once inserted, lexical function names are non-editable text.

4. Conclusion

As mentioned at the very beginning of this paper, we are presenting a new lexicographic project and it is too early for us, at the time of writing, to be able to draw any conclusion from our theoretical and methodological choices. However, we believe the content, structure and methodological design of the FLN to be original enough to generate interest for anyone concerned with the construction and availability of multi-purpose lexical resources. Of particular relevance is the fact that the FLN is designed as a truly generic database. It targets NLP exploitation—that imposes very strong formal constraints on lexical data—as well as pedagogical exploitation—that shows zero tolerance to error in the modeling of linguistic rules.

Note that the FLN will be made available on the CNRTL website[19] in the course of its growth, both as a source SQL database and via a web-based interface for manual consultation. It is also our intention to later explore the possibility to generate LMF[20] compatible exports of FLN data.

[19] http://www.cnrtl.fr/

[20] Lexical Markup Framework, ISO-24613:2008 (Francopoulo et al., 2009).

Acknowledgements

Many thanks to Sébastien Haton, Jasmina Milićević, Dorota Sikora and two reviewers for WoLeR 2011 for their comments on a preliminary version of this paper. We are very grateful to Pascale Lefrançois (Université de Montréal) and Ophélie Tremblay (Université du Québec à Montréal) for giving us access to their research material, that helped us greatly in the construction of the FLN priming wordlist; we additionally thank Caroline Bégin (MELS) and Hélène Cajolet-Laganière (Université de Sherbrooke) for authorizing the dissemination of the information contained in Lefrançois et al. (2011). The RELIEF project is supported by a grant from the Agence de Mobilisation Économique de Lorraine (AMEL) and Fonds Européen de Développement Régional (FEDER).

5. References

J. Aitchison. 2003. *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell, Oxford UK, 3rd edition.

C. F. Baker, C. J. Fillmore, and B. Cronin. 2003. The Structure of the FrameNet Database. *International Journal of Lexicography*, 16(3):281–296.

L. Barque, A. Nasr, and A. Polguère. 2010. From the Definitions of the *Trésor de la Langue Française* To a Semantic Database of the French Language. In A. Dykstra and T. Schoonheim, editors, *Proceedings of the XIV Euralex*

International Congress, pages 245–252, Leeuwarden, 6–10 July. Fryske Akademy.

M. Benson, E. Benson, and R. Ilson. 1997. *The BBI Dictionary of English Word Combinations*. John Benjamins, Amsterdam/Philadelphia, revised edition.

Collectif Robert. 2009. *Le Robert Benjamin*. Le Robert, Paris.

G.-M. de Schryver and G. de Pauw. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of TshwaneLex. *Lexikos*, 17:226–246.

J. Dendien and J.-M. Pierrel. 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.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press., Cambridge, MA.

T. Fontenelle. 1997. *Turning a bilingual dictionary into a lexical-semantic database*. Niemeyer, Tübingen.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2009. Multilingual Resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70.

G. Gougenheim, R. Michéa, P. Rivenc, and A. Sauvageot. 1967. *L'élaboration du français fondamental*. Didier, Paris.

F. J. Hausmann. 1979. Un dictionnaire des collocations est-il possible ? *Travaux de littérature et de linguistique de l'Unviersité de Strasbourg*, XVII(1):187–195.

D. Hirsh and P. Nation. 1992. What Vocabulary Size Is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language*, 8(2):689–696.

A.-L. Jousse and A. Polguère. 2005. *Le DiCo et sa version DiCouèbe. Document descriptif et manuel d'utilisation*. Technical report, Department of Linguistics and Translation, Université de Montréal.

A.-L. Jousse. 2010. *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. Ph.D. thesis, Université de Montréal & Université Paris Diderot (Paris 7), Montreal & Paris.

S. Kahane and A. Polguère. 2001. Formal Foundation of Lexical Functions. In *Proceedings of the Workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation"*, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics, pages 8–15, Toulouse, 7 July 2001.

F. Lareau. 2002. A practical guide to writing *DiCo* entries. In *Proceedings of PAPILLON 2002 International Workshop on Multilingual Lexical Databases*, Tokyo, 16–18 July.

P. Lefrançois, O. Tremblay, and V. Lombard. 2011. Constitution de listes de mots pour l'apprentissage de l'orthographe et du lexique au primaire et au début du secondaire. Research report for the *Ministère de l'Éducation, du Loisir et du Sport du Québec (MELS)*, Université de Montréal, Montreal, 14 February 2011.

I. Mel'čuk and A. Polguère. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Champs linguistiques. De Boeck & Larcier, Brussels.

I. Mel'čuk, A. Clas, and A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Paris/Louvain-la-Neuve.

I. Mel'čuk et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques. Volumes I–IV*. Les Presses de l'Université de Montréal, Montréal.

I. Mel'čuk. 1995. Phrasemes in Language and Phraseology in Linguistics. In Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 167–232. Laurence Erlbaum Associates, Hillsdale, N.J.–Hove, UK.

I. Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins, Amsterdam/Philadelphia.

I. Mel'čuk. 2006. Explanatory Combinatorial Dictionary. In Giandomenico Sica, editor, *Open Problems in Linguistics and Lexicography*, pages 225–355. Polimetrica, Monza.

P. Nation. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, 63(1):59–81, September.

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

A. Polguère. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43(1):41–55, March. Springer.

A. Polguère. To appear. Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de lexicologie*.

J. Rey-Debove and A. Rey, editors. 2010. *Nouveau Petit Robert*. Le Robert, Paris.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley CA.

T. Selva, S. Verlinde, and J. Binon. 2003. Vers une deuxième génération de dictionnaires électroniques. *Traitement Automatique des Langues (T.A.L.)*, 44(2):177–197.

J. Steinlin, S. Kahane, and A. Polguère. 2005. Compiling a "classical" explanatory combinatorial lexicographic description into a relational database. In *Proceedings of the Second International Conference on the Meaning Text Theory (MTT'2005)*, pages 477–485, Moscow.

D. Summers, editor. 2005. *Longman Dicionary of Contemporary English*. Pearson Longman, Essex, 4th edition.

F. Ters, G. Mayers, and D. Reichenbach. 1988. *L'échelle Dubois-Buyse*. OCDL, Paris.