Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor

Nabil GADER[1] Veronika LUX-POGODALLA[2] Alain POLGUÈRE[3]
(1) MVS Publishing Solutions, Sainte-Marguerite, F-88100, FRANCE
(2) CNRS, ATILF, UMR 7118, Nancy, F-54063, FRANCE
(3) Université de Lorraine, ATILF, UMR 7118, Nancy, F-54000, FRANCE
Nabil.Gader@mvs.fr, Veronika.Lux@atilf.fr,
Alain.Polguere@univ-lorraine.fr

ABSTRACT
We present the data structure of a lexical resource—the French Lexical Network (FLN)—, that is being hand-crafted using a knowledge-based lexicographic editor. The FLN is formally a lexical graph whose structuring is mainly supported by the system of paradigmatic and syntagmatic lexical functions of the Meaning-Text linguistic approach. Section 1 offers a general characterization of the FLN. Section 2 describes the database and the lexicographic editor in their present state. Section 3 focuses on the SQL data structure used to encode lexical information in the FLN, with special attention paid to the encoding of lexical function relations. Section 4 considers the feasibility of porting the FLN data to known standards such as the Lexical Markup Framework (LMF). Finally, in section 5, we consider the cognitive relevance of the FLN approach to the modeling of lexicons.

KEYWORDS: lexical database, lexical graph, virtual dictionary, semantic derivation, collocation, lexical function, Explanatory Combinatorial Lexicology/Lexicography, lexicographic editor, French language.

Introduction

The *French Lexical Network*, hereafter FLN,[1] is a new hand-crafted lexical resource, currently under development, that possesses many distinguishing features, both in terms of content, structure and building process. In this paper, we focus on the FLN's data structure and on the graph editor that has been designed to support the lexicographic task of building the FLN. This work is currently performed at the ATILF CNRS laboratory (Nancy, France) in the context of a broader R&D project called *RELIEF* (Lux-Pogodalla and Polguère, 2011). Though the process of building the FLN is a long-term enterprise and we are at the time of writing only 18 months into this project, the resource's structure is already sufficiently stable, and the resource itself is sufficiently well into development, for us to be able to account for our first results. We believe that the approach taken in designing the FLN is particularly relevant for the linguistics, NLP and cognitive science communities due to (i) its formal nature, (ii) its strong theoretical linguistic background and (iii) its fundamental semantic orientation. All points that will be made clearer below.

[1] In French: *Réseau Lexical du Français* or *RLF*.

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 109–126,
COLING 2012, Mumbai, December 2012.

1 General characterization of the French Lexical Network (FLN)

The FLN belongs to the family of Net-like lexical databases (Fellbaum, 1998; Baker et al., 2003; Ruppenhofer et al., 2010; Spohr, 2012) and possesses at least four distinguishing characteristics.

1. The FLN is a lexical network—i.e., a network of interconnected lexical units—whose structure is mainly organized around a constantly growing set of lexical links, based on the system of so-called *lexical functions* proposed by the Meaning-Text linguistic theory.[2]

2. Though manually performed, the construction of the FLN is done by means and "under the supervision" of a tailor-made lexicographic editor named *Dicet*—developed by MVS Publishing Solutions (Sainte-Marguerite, France)[3]—that allows lexicographers to browse through the lexical network and directly expand and revise it, using linguistic and metalinguistic information. Dicet can therefore be best conceived of as being a knowledge-based lexical graph editor and browser.

3. Lexical information stored in the FLN is entirely formalized, thus allowing for computer processing of the lexical network, for both lexicographic purposes (automatic coherence checking, implementation of analogical lexicographic reasoning, etc.) and natural language processing.

4. Though not a dictionary—it doesn't possess the textual structure of a paper or computerized dictionary—the FLN is designed to have embedded in it sufficient lexicographic information (both in formal and "popularized" form) to be used to automatically generate dictionaries of multiple formats. It is thus the repository of *virtual dictionaries* (Atkins, 1996; Selva et al., 2003; Polguère, 2012a).

This last characteristic is particularly important as it explains why the construction of the FLN is indeed a true lexicographic project. Ultimately, the FLN is meant to be a multi-purpose lexical resource, that should allow for the automatic generation of (i) dictionaries with various macro- and microstructures targetting human users, and (ii) formalized resources for NLP. The fact that computer programs should be able to make use of the FLN sets the target in terms of formalization and "computability." On the other hand, the targeting of a content that is dictionary-grade and suitable for human users—language learners being the prototypical users—sets very high standards in terms of accuracy. This rules out any strategy of automatically compiling the FLN out of already existing linguistic resources such as electronic dictionaries, corpora, etc. (Sagot and Fišer, 2011). The process of building the FLN has to be a full-fledged lexicographic one and involves an organized team of lexicographers.[4]

2 Present state of the lexical graph and lexicographic editor

Let us first mention that, according to our terminology, a *lexical entry* in the FLN corresponds to a (potentially) polysemic word, called *vocable*. We name *lexical unit* each well-specified sense of the vocables that constitute the FLN's wordlist. For instance, the French vocable CHANTAGE

[2] See (Milićević, 2006) for a short introduction to the Meaning-Text approach to language study.

[3] MVS is ATILF's private sector partner in the RELIEF project, that has a significant R&D facet. RELIEF targets both the construction of the FLN and its utilization in natural language processing tasks such as fine-grained semantic access to textual information.

[4] At present, 12 members of the team are directly involved in lexicographic tasks.

'blackmail' comprises two senses—i.e. lexical units—in the current state of the FLN: CHANTAGE **I** 'criminal act' and CHANTAGE **II** 'pressure (put on someone).'

Each lexical unit possesses a unique ID (identifier) in the FLN's database. Such is the case for each vocable, the belonging of lexical units to given vocables being modeled as relations between vocable and lexical unit IDs.

Links between lexical units are also implemented as links between lexical unit IDs. This is true in particular for links based on so-called *lexical functions* of the Meaning-Text linguistic approach (Mel'čuk, 1996), that form the bulk of the FLN's structure. For instance, let us return to the case of Fr. CHANTAGE **I**, whose predicative structure is 'blackmail by $1 on $2 regarding $3 to obtain $4.'[5] The fact that Fr. CIBLE **II.2** 'target' and VICTIME **II** 'victim' are typical names for the second actant ($2) of CHANTAGE **I** is modeled by the following *lexical function application*, where $\mathbf{S_2}$ is the paradigmatic lexical function that returns typical names for the second actant of a predicative lexical unit:

$$\mathbf{S_2}(\textit{chantage}\,\mathbf{I}) = \textit{cible}\,\mathbf{II.2}\ [\textit{de}\ \text{ART}\sim],\ \textit{victime}\,\mathbf{II}\ [\textit{de}\ \text{ART}\sim]$$

Such lexicographic information about CHANTAGE **I** is structured in the FLN by means of the following lexical subgraph:

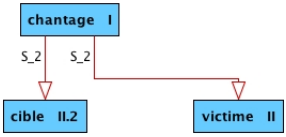

Figure 1: Structure of $\mathbf{S_2}$(*chantage* **I**) in the FLN

Paradigmatic lexical functions, such as $\mathbf{S_2}$, correspond to semantic relations between lexical units called *semantic derivations* in Meaning-Text linguistics terminology. However, there exist also syntagmatic lexical functions, that correspond to collocational links between lexical units. For instance, the fact that the verbs Fr. CÉDER **IV.1** lit. 'to give in [to someone]', OBÉIR **3** lit. 'to obey' and OBTEMPÉRER lit. 'to comply' are used as collocates of their complement CHANTAGE **I** to express '$2 does what he/she is expected (by $1) to do in respect to $1's blackmail,' is modeled by the following lexical function application, where $\mathbf{Real_2}$ is the syntagmatic lexical function that returns "verbs of realization" that take the second actant of a predicative noun as subject and the noun itself as first complement:

$$\mathbf{Real_2}(\textit{chantage}\,\mathbf{I}) = \textit{céder}\,\mathbf{IV.1}\ [\textit{à}\ \text{ART}\sim],\ \textit{obéir}\,\mathbf{3}\ [\textit{à}\ \text{ART}\sim],\ \textit{obtempérer}\ [\textit{à}\ \text{ART}\sim]$$

In total, the set of lexical function links that gravitates around a given lexical unit can be quite significant. This is illustrated in Figure 2 below, that displays all lexical function links of which

[5] $1, $2, $3,... are local variables (in the computational sense) that function locally in each individual lexical unit description. They ensure the proper numbering and naming of actant slots and are used in place of the traditional *X*, *Y*, Z, ... variables.

CHANTAGE I is currently the source or the target.[6] Notice that this lexical unit represents a mild case in terms of lexical function connections: many links leaving from and leading to CHANTAGE I are yet to be encoded and it is easy to find lexical units that are much bigger "lexical crossroads" than this particular one.

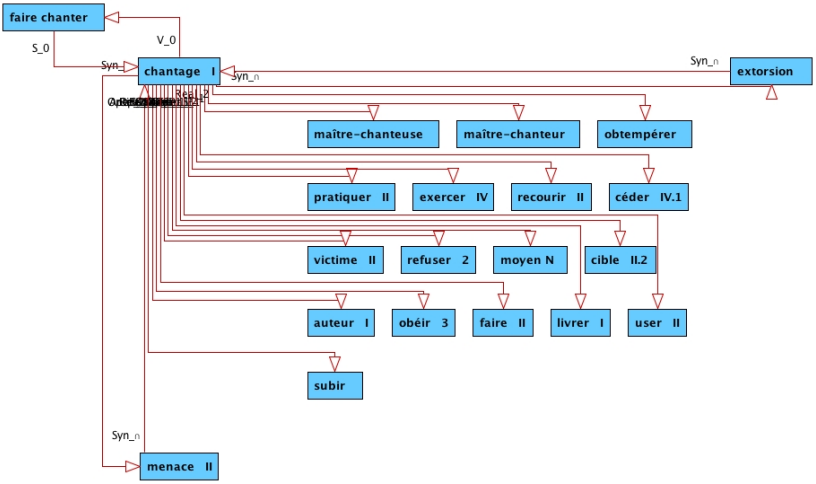

Figure 2: All lexical function links leaving from or leading to CHANTAGE I in the FLN

As one can see, the bulk of the FLN's informational content and structuring lies in the set of lexical units (grouped under given polysemic vocables) and the set of lexical function links that connect lexical units together. Based on the formal characteristics of the resource, the best way to evaluate the FLN's state of development is to count:

1. the number of lexical entities—mainly, vocables (V) and lexical units (LU)—it contains, such as for standard dictionaries where the number of entries and senses are often used as coverage measurement;
2. the polysemy rate (LU/V), that tells us how many lexical units (senses), in average, are grouped under each vocable;
3. the number of lexical function links (LFL) between lexical units;
4. the connectivity rate LFL/LU, that tells us how many lexical function (in or out) links are connecting, in average, a lexical unit to the rest of the FLN graph.[7]

[6] For lack of space we are forced to select a display that doesn't allow for the legibility of all arc labels. All lexical graphs used here have been automatically generated from the FLN database and displayed by means of the yEd graph editor (`http://www.yworks.com/en/products_yed_about.html`).

[7] Expressed in mathematical terms (graph theory), the connectivity rate is the average degree of lexical nodes in the FLN graph whose edges are lexical function relations.

Here are the statistics at the time of writing:

Vocables, i.e. entries [$= V$]	: 10363
Lexical units, i.e. senses [$= LU$]	: 13767
Polysemy rate [$= LU/V$]	: 1.33
Lexical function links $LU_1 \rightarrow LU_2$ [$= LFL$]	: 17353
Connectivity rate [$= LFL/LU$]	: 1.26

As the first two numbers show, the FLN is already far from being a sample or prototype of a lexical resource in regard to how many entries and senses it contains. In actual fact, we do consider that we have already reached our target in the RELIEF project in terms of wordlist size. We initially estimated the minimal number of entries to be around 10,000, with no fixed upper limit. However, adding an entry to the RLF wordlist is not a difficult task once a core wordlist for French has been identified. What matters from now on is the growth of information that is to be attached to each vocable: identification of its polysemy through creation of senses associated to it and description of linguistic properties of each senses—essentially, through weaving of lexical function links.

The polysemy rate is a particularly good indicator of how advanced the description of each vocable is. Each time a vocable is actually studied and described, its sense structure is analyzed and described by adding new senses to the database. A good polysemy rate for a fully mature database would be between 2.5 and 3. For the sake of comparison, the French *Petit Robert* reference dictionary—a very detailed dictionary in respect to the polysemic structuring of entries—possesses a polysemy rate of 5. The rate of 1.33 that we currently achieve indicates that the FLN has not reach its full maturity yet, but is already an "adolescent" lexical database on its way to adulthood.

> Let us emphasize the fact that statistics given here are based on the complete database, not on lexical units that have actually been methodically studied. All rates are therefore "diluted" by the mass of targeted units that are participating in holding the graph together but are still unexplored locations in the global RLF topography.

Together with the polysemy rate, the connectivity rate is an important indicator of how fleshy and informative each entry is. Figure 3 below indicates the evolution of statistics on connectivity since the moment the hardcoding of lexical function relations has been launched.

Notice that, at the moment of its birth, the FLN was nothing but a "fully non-connected" graph: a set of individual nodes (lexical units) with no lexical connection to other nodes. This initial set of 3,734 nodes was automatically injected into the RLF database from a manually constructed *priming wordlist*—see (Lux-Pogodalla and Polguère, 2011; Polguère and Sikora, ToAp) for details on the FLN's growth process.

As for the FLN microstructure, lexical units—*headwords*—articles are made up of seven lexicographic zones:

1. `GC` for grammatical characteristics (part of speech, noun gender, specific inflectional behavior, etc.);
2. `DF` for definition;

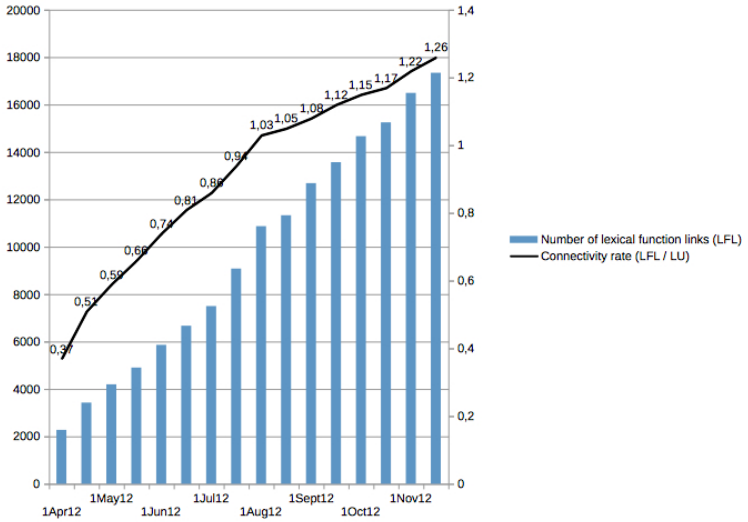

Figure 3: Evolution of lexical connectivity in the FLN

3. NB for *Nota Bene* about the headword description;
4. GP for the headword's government pattern, i.e. the description of its syntactic valency (Mel'čuk, 2004a,b; Milićević, 2009);
5. LF for lexical function relations originating from the headword (headword as argument of lexical function applications);
6. EX for lexicographic exemples;
7. PH for pointers to so-called *full phrasemes*, i.e. idioms, that are formally made up of a lexemic headword—e.g. BULLET points to BITE THE BULLET.

All lexicographic zones are currently being dealt with by lexicographers. However, only the GC (grammatical characteristics) and LF (lexical functions) zones are fully formalized and supervised by the Dicet editor at the time of writing. The other zones are for the time being completed as simple text fields and the EX (lexicographic examples) zone is presently under formalization and about to be completed at the time of writing.

By saying that a given zone is supervised by the Dicet editor, we mean that:

- Dicet possesses knowledge about the information that has to be provided in the zone;
- embedded in Dicet, are special lexicographic tools that allow for the entering of information under complete supervision of the editor.

In this approach to lexicography, lexicographers do not *write* articles; rather, they *build* them by putting together all microscopic lexical rules that are associated with each lexical unit. The encoded information is used by the editor for computing a textual presentation in what is called an *article-view* of the headword's description. At no time does the lexicographer type a lexicographic text in an implemented zone (except in its `Comments` field). Figure 4 below shows an association between the sub-window of the `LF` zone that is used to pull lexical function links from the headword—at the bottom, right above the `Comments` zone—and the corresponding article-view—on top—that displays the encoded information in textual (dictionary-like) form. The headword used in this figure is CHANTAGE **I**, whose position in a lexical function subgraph of the FLN has been described above (see Figure 2).

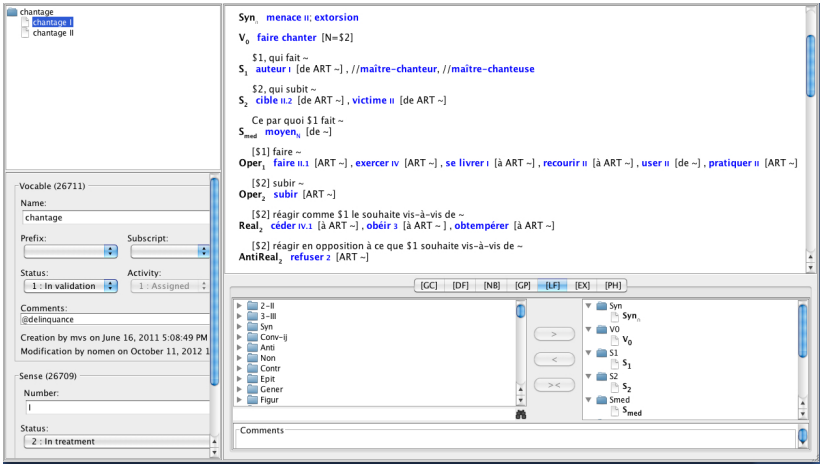

Figure 4: Correspondance between lexical function lexicographic tools and the article-view

Names of targets of lexical function links in the article-view are clickable textual items that give direct access to the edition of the corresponding lexical units (opening of new editing windows). For lack of space, we cannot delve further into the functioning of the Dicet editor. But we will have the opportunity to provide more information on its functionalities when presenting the FLN's computational model (next section).

The designing and programming of the Dicet editor represents a significant investment in terms of time and effort; before undertaking it, we reviewed existing softwares without being able to find anything that came close to what we were looking for: a lexicographic editor supporting graph weaving of lexical function links. It is also important to highlight the fact that Dicet was not built from scratch. It is a customization of resources that were already available from the MVS partner of the project, among its publishing solutions: primarily, the *Dixit* publishing tool.[8] Dicet is indeed a lexicographically-boosted knowledge-based version of Dixit, recoded in Java for portability purposes. Dicet is also part of a suite integrating workflow, user rights

[8] `http://www.mvs.fr/pdf/MVS_Dixit.pdf`

management and controlled connection to an SQL database, while most existing lexicographic editors seem to work on XML databases— for example, IDM DPS (Lannoy, 2010) and the Dictionary Editor and Browser (DEB) (Horák et al., 2006).

In spite of its distinctive features, Dicet shares features with tools such as TLex (Joffe and de Schryver, 2012) or the above-mentioned DEB. In particular, just as TLex supports "smart cross-references" (Joffe and de Schryver, 2012, pp. 25–27 and pp. 68–69), Dicet allows easy creation and maintenance of links between lexical units, a major concern given our lexicographic model. However, beyond lexicographic editing, Dicet was designed to support a completely new approach to building lexical resources. Because we developed our own tool, we were totally free to explore and implement new ways of performing lexicographic activity (cf. section 5 below).

3 Data model: a relational SQL database

The FLN's data model is an SQL database which, at the time of writing, comprises 46 separate SQL tables; presenting all of them here is of course out of the question. As very limited space is available to us, we will concentrate on lexical functions. Notice that a significant number of publications on formal and computational modeling of lexical functions are already available, for instance (Kahane and Polguère, 2001; Iordanskaja et al., 1992; Lareau et al., 2012); we will consider solely the FLN's approach to the problem.

In total, 16 SQL tables are used in the FLN's database for the storage of lexical function-related information. Part of these tables are used for modeling lexical functions per se, e.g. $\mathbf{S_2}$, as individual lexical entities; this information could be exported as a stand-alone model of the linguistic system of lexical functions (cf. section 4). Figure 5 below shows the interface that allows lexicographers to manage the lexical function knowledge base embedded in the FLN. In this figure, one can see the database record that defines the $\mathbf{S_2}$ lexical function. It contains the three following types of information, from top to bottom.

1. Classification: $\mathbf{S_2}$ is a simple standard paradigmatic lexical function of the "$\mathbf{S_2}$" family, that comprises also $\mathbf{S_{2\supset}}$, $\mathbf{S_{2\cap}}$, $\mathbf{S_2^{usual}}$, etc.
2. Formula structure: the lexical function formula $\boldsymbol{S_2}$ is constructed by assembling two atomic formal elements—the name central component $\boldsymbol{S}$ and the subscript $\boldsymbol{2}$. As one can see, names of lexical functions are broken down into atomic building blocks[9], thus allowing for future automatic compilation of standard lexical function encodings into more "computable" formulas, such as those proposed in (Kahane and Polguère, 2001).
3. Popularization: $\mathbf{S_2}$ is at present associated with three popularization formulas (from the popularization formula database).

In the remainder of this section, we focus on the SQL modeling of lexical function *applications*— e.g. $\mathbf{S_2}$(*chantage* I)—rather than lexical function themselves. This modeling is supported by 4 SQL tables, shown in Figure 6 below. This subpart of the database can be seen as storing actual lexical function relations among lexical units. As said earlier, such data represents the crucial element of lexical structuring in the FLN.

The content of each of the 4 SQL tables in Figure 6 can be described as follows.

[9]For instance, the complex lexical function formula $\boldsymbol{Magn^{quant}+A_1}$ is defined in the FLN database as an assembling of 5 atomic elements: $\boldsymbol{Magn}$, $\boldsymbol{^{quant}}$, $+$, $\boldsymbol{A}$ and $\boldsymbol{_1}$.

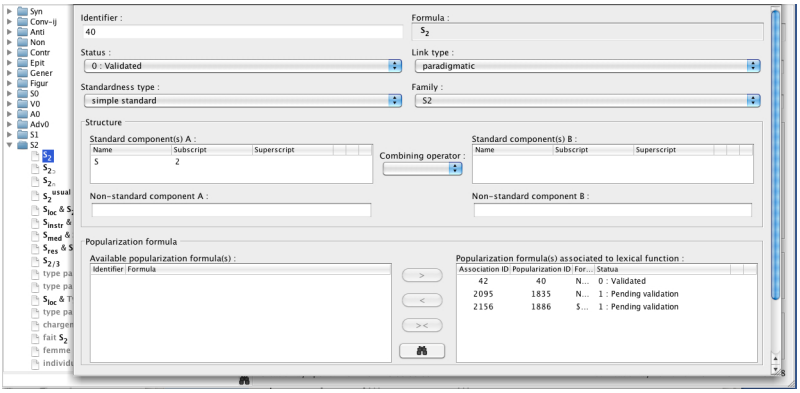

Figure 5: Interface for managing the FLN lexical function knowledge base: definition of $\mathbf{S_2}$

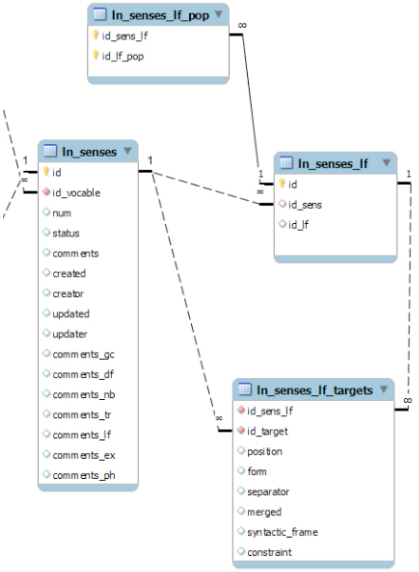

Figure 6: SQL tables handling lexical function applications in the FLN

1. Table **ln_senses** describes each lexical unit (sense), using 16 different fields. A lexical unit, identified by a unique identifier (id), is formally linked to the vocable it belongs to (based on id_vocable) and is characterized by a lexicographic number (ex. **I.1**) if its vocable is polysemic.[10] To sum up, the ln_senses table is used to store basic information about lexical units such as CHANTAGE **I**, CIBLE **II.2**, VICTIME **II**, CÉDER **IV.1**, OBÉIR **3**, OBTEMPÉRER, etc.

2. Table **ln_senses_lf** describes the application of a lexical function **LF** (whose unique identifier is id_lf) to a lexical unit L (whose unique identifier is id_sens). Each application of a lexical function to a lexical unit **LF**(L) has a unique identifier (id in table ln_senses_lf).

3. This identifier is used under the name id_sens_lf in the **ln_senses_lf_pop** table, that handles the association between individual lexical function applications and the associated popularization formula. For example, the table ln_senses_lf is used to store:

   - the application of the lexical function $\mathbf{S_2}$ to the lexical unit CHANTAGE **I**, application to which the table ln_senses_lf_pop associates the popularization formula [$2] qui subit ~ (= '[$2] who undergoes ~').[11]
   - the application of the lexical function $\mathbf{Real_2}$ to the lexical unit CHANTAGE **I**, application to which the table ln_senses_lf_pop associates the popularization formula [$2] réagir comme $1 le souhaite vis-à-vis de ~ (= '[$2] to react as expected by $1 regarding ~').

4. Each target L' of a lexical function application **LF**(L) is specified in the **ln_senses_lf_targets** table, using lexical function identifiers. But the ln_senses_lf_targets table also contains:

   - information necessary to logically order all **LF**(L)'s targets (field position) when they are enumerated in an article-view;
   - information about target separators (field separator whose value can be "**,** ", "**;** " or "<");
   - information about the complementation frame of each target (field syntactic_frame);
   - etc.

   In short, this table stores all additional linguistic information that is necessary to compute what is displayed in the article-view for the lexical function zone (see the article-view in Figure 4 above).

[10] Some fields, whose names begin by comment_, are used for the storage of "freely" entered information, i.e. information for which Dicet does not yet implement a full formalization (comments_gc stands for *comments on grammatical characteristics*, comments_df stands for *comments on definition*, etc.). Some other fields are for the management of the lexicographic work. E.g. status has a value indicating if the description of the lexical unit is "completed," "being validated," "under description" or "unworked;" created contains the creation date; creator contains the login of the sense's creator; etc.

[11] The "~" symbol is used throughout a lexicographic article to refer to this article's headword.

When weaving lexical function relations among lexical units, lexicographers work under the supervision of the Dicet editor. Figure 7 shows the interface they use to feed the `ln_senses_lf_targets` table with all the necessary information.

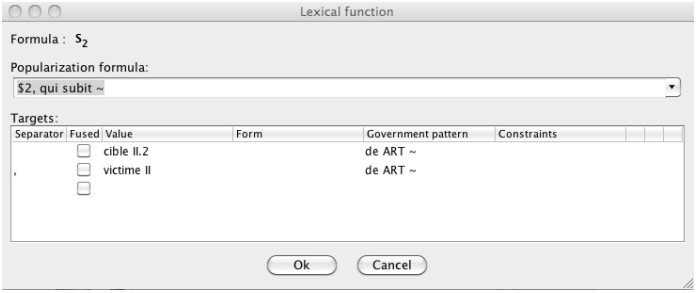

Figure 7: Part of the Dicet interface for pulling lexical function links in the FLN

Because our lexicographic model is implemented as SQL tables and the FLN itself is stored in a relational database, we benefit from the well-established technology of relational database management systems (quick access to data, secure data storage, management of users rights, etc.).

As illustrated in this section, the Dicet editor is a powerful interface. Dicet helps lexicographers to enter data that is compliant to our lexicographic principles. Simultaneously, Dicet computes article-views of lexicographic data that is instantly available to lexicographers as a retroaction they can use in order to check and validate their description. In short, Dicet ensures compliance with a fine-grained formal model while giving lexicographers the clear feeling that they are, afterall, performing a task that is equivalent to "writing" lexicographic articles.

4 Compatibility with standards for lexical data structures

It is nowadays unconceivable to target the public distribution of a lexical resource such as the FLN without taking into consideration the compatibility of its computational modeling of linguistic information (i.e., data structure) with available standards. In this section, we present the outcome of some preliminary reflections on this topic.

As explained above, the FLN is implemented as a relational SQL database. This database provides XML (and HTML) data exports, which means that it is conformant with a general well-known data model and a general format standard. We wish to explore the compatibility of FLN's data with a more specialized standard: the *Lexical Markup Framework*—hereafter, LMF—, which is the ISO standard for NLP-compatible lexical databases and dictionaries (ISO, 2008; Francopoulo et al., 2006). LMF Core Package and extensions are defined with the Unified Modeling Language (UML), so that the normative content of LMF is expressed as sets of UML classes with associations among classes. Attribute-value pairs used to adorn the UML classes are not directly provided by LMF. Rather, LMF recomands to use Data Category specifications in accordance with the ISO 12620 standard (ISO, 2009).

To start with, we have checked if the FLN is in accordance with LMF general principles expressed in the LMF core package. The `Lexical Entry` class from the LMF Core Package, which is

just “a container for managing the Form and Sense classes” (ISO, 2008, p. 18) perfectly matches the FLN vocable entity, which we model as a grouping of lexical units. More importantly, in both LMF and the FLN, the basic unit for lexicographic description is a `Sense` (= lexical unit in the FLN).

To go further, since the FLN provides a large range of properties for each lexical unit, we would have to select several LMF extensions in addition to the LMF core package. We chose to focus on the encoding of lexical function applications—whose central role in structuring the FLN has been extensively discussed above—and examine how it can be compiled into LMF format.

The LMF NLP semantics extension includes a `SenseRelation` class, defined as “a multipurpose class that can be used to represent *antonymy*, *generic/specific* or *part of* relationship” (ISO, 2008, p. 41), which seems to fit our needs. The code sample below illustrates the use of this class to model the two lexical function applications $\mathbf{S_2}$(*chantage*I) and $\mathbf{Real_2}$(*chantage*I), that have been examined in section 2.

```
<LexicalEntry>
    <feat att="partOfSpeech" val="nom"/>
    <Lemma>
        <feat att="writtenForm" val="chantage"/>
    </Lemma>
    <Sense xml:id="chantage:I">
        <SenseRelation targets="#cible:II.2 #victime:II">
            <feat att="lexicalFunction" val="S_2"/>
            <feat att="popularizedFormat" val="[$2] subir ~"/>
        </SenseRelation>
        <SenseRelation targets="#céder:IV.1 #obéir:3 #obtempérer">
            <feat att="lexicalFunction" val="Real_2"/>
            <feat att="popularizedFormat" val="[$2] réagir
                  comme $1 le souhaite vis-à-vis de ~"/>
        </SenseRelation>
    </Sense>
</LexicalEntry>
<LexicalEntry>
    <feat att="partOfSpeech" val="nom"/>
    <Lemma>
        <feat att="writtenForm" val="cible"/>
    </Lemma>
    <Sense xml:id="cible:I"/>
    <Sense xml:id="cible:II.1"/>
    <Sense xml:id="cible:II.2"/>
</LexicalEntry>
<!-- Lexical entry omitted for "victime" -->
<LexicalEntry>
    <feat att="partOfSpeech" val="verbe"/>
    <Lemma>
        <feat att="writtenForm" val="céder"/>
    </Lemma>
    <Sense xml:id="céder:I"/>
    <Sense xml:id="céder:II"/>
    <Sense xml:id="céder:III"/>
    <Sense xml:id="céder:IV.1"/>
    <Sense xml:id="céder:IV.2"/>
</LexicalEntry>
<!-- Lexical entries omitted for "obéir" and "obtempérer" -->
```

The above code is based on the LMF `SenseRelation` class as it is, using only one additional

feature to encode the associated popularization formula. For an actual compilation of the FLN into LMF, we will probably need to define a new class, say `LexicalFunctionRelation`, as a specialization of the `SenseRelation` class.

Additionally, this example clearly shows that the XML version of the LMF model, used here, is not rich enough. In particular, the `targets` attribute of the `SenseRelation` element allows us to link a lexical unit to a list of lexical units, using a given lexical function, while, as shown in section 3, we need a data structure that is far more complex than a list, together with additional information attached to links pointing to targeted lexical units.[12] We need a true LMF equivalent of our `ln_senses_lf_targets` SQL table.

From our preliminary exploration of the problem of the FLN's compatibility with existing standards, we draw the following conclusions.

- The FLN is in accordance with the few high-level good-practice principles provided by LMF.
- Since the FLN has a very precisely defined structure, we need to add more specificity to LMF classes in order to get a tighter LMF model.
- If an LMF compatible distribution of the FLN is indeed implemented, several linguistic resources used in the FLN should be converted into Data Category Registries, following the ISO 12620 standard (ISO, 2009). For example, in order to define the attributes and values used to adorn the `LexicalFunction` class, one probably has to build a Data Category Registry dedicated to lexical functions. This could be automatically generated from a subset of the FLN SQL tables, in which a description of all FLN's lexical functions is provided.
- Finally, let's mention that we consider exploring in the future other standards for lexical resources, in particular the Text Encoding Initiative (TEI), that includes two relevant chapters: *Dictionaries* and *Graphs, networks and trees*.

5 On the cognitive relevance of the FLN approach

In order to conclude, we wish to reflect on the cognitive relevance of the FLN approach to the modeling of lexical information. Indeed, our motivations for designing such a project—in terms of lexical model and lexicographic methodology—do not originate from computational considerations. The need to implement, fully formalize and make computer-tracktable our lexical model is a non-negotiable constraint, an essential parameter in our approach. However, our first and foremost goal is to build a lexical resource that complies before all not to encoding standards, but to lexicological ones! The FLN design and modus operandi is the outcome of an extremely long process of experimentation with lexicological models and of lexicographic practice: from earlier "theoretical dictionaries" called *Explanatory Combinatorial Dictionaries* (Mel'čuk and Žolkovskij, 1984; Mel'čuk, I. *et al.*, 1999), to the work on the *DiCo* lexical database (Polguère, 2000; Mel'čuk and Polguère, 2006), the layman-oriented pedagogical *Lexique Actif du Français*[13] (Mel'čuk and Polguère, 2007; Polguère, 2007) and the first proposal for a graph-based version of Explanatory Combinatorial lexical databases called *lexical systems* (Polguère,

[12] Cf. Figure 7, section 3, the set of parameters represented by columns to be filled in the Dicet window.
[13] Lit. 'Active French Lexicon.'

2009). All this has matured into a project that, we hope, goes beyond the very specific problem of building a French lexical resource.[14]

As was mentioned at the beginning of this paper (section 1), the constraint of being able to use the FLN as a resource for such a linguistically demanding context of application as language learning (and teaching) imposes on us to consider a lexical model that has relevance to the processes of acquiring and using lexical knowledge. There are at least two aspects of the FLN that we believe make it compatible with this goal.

Firstly, the FLN is a rich, non-hierarchical lexical graph, that is more in line with the plausible structure of "actual" lexical knowledge (Aitchison, 2003) than textual models of the dictionary type.

Secondly, the Dicet editor has a crucial importance in our approach in that it not only helps entering and retrieving formally coherent information, but it also implements a new "lexicographic gesture." We are convinced that this gesture is intrinsically compatible with language speaker's navigation in the lexicon in the context of language learning and use (Wolter, 2006; Zock and Schwab, 2011). We cannot enter here into the detail of this last aspect of our work; it is dealt with in A. Polguère's oral presentation at CogALex III (Polguère, 2012b) and will be developed in later publications. Suffice it to say here that FLN's lexicographers build the lexical model in a non-linear way, through gradual and sometimes aleatory weaving of lexical links. This process of building lexicographic information—that follows semantic, combinatorial and formal relations between lexical units—presents strong analogies with plausible wading of the speaker through the structure of lexical knowledge.

Acknowledgments

The RELIEF project is supported by a grant from the Agence de Mobilisation Économique de Lorraine (AMEL) and Fonds Européen de Développement Régional (FEDER). We wish to thank our colleague Bertrand Gaiffe for his precious guidance on LMF and CogALex III reviewers for their extremely sound and useful comments on a preliminary version of this paper.

References

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

Atkins, B. T. S. (1996). Bilingual Dictionaries: Past, Present and Future. In Gellerstam, M., Järborg, J., Malmgren, S.-G., Norén, K., Rogström, L., and Papmehl, C. R., editors, *Euralex'96 Proceedings*, pages 515–590, Gothenburg. Gothenburg University, Department of Swedish.

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

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

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of International Conference on Language Resources and Evaluation – LREC 2006*, Genova.

[14]Note that the FLN approach is presently being used, in exploratory satellite projects based on the same data structure and lexicographic editor, for the modeling of the Korean and Spanish lexicons.

Horák, A., Pala, K., Rambousek, A., and Rychlý, P. (2006). New Clients for Dictionary Writing on the DEB Platform. In de Schryver, G.-M., editor, *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems*, pages 17–23, Turin.

Iordanskaja, L., Kim, M., and Polguère, A. (1992). Some Procedural Problems in the Implementation of Lexical Functions. In Wanner, K. H. . L., editor, *Proceedings of the International Workshop on The Meaning-Text Theory*, Arbeitspapiere der GMD 671, pages 197–205, Darmstadt (Allemagne). GMD-MBH.

ISO (2008). Language Resource management – Lexical markup framework (LMF). ISO/TC 37/SC 4 N453. N330 Rev. 16.

ISO (2009). Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources. ISO/TC37/SC3 ISO 12620. Stage : 60.60 (2009-12-10).

Joffe, D. and de Schryver, G.-M. (2012). TLex Suite User Guide (version 7.0.1). Technical report, TshwaneDJe Human Language Technology.

Kahane, S. and Polguère, A. (2001). Formal Foundation of Lexical Functions. In *Proceedings of "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.

Lannoy, V. (2010). The IDM Free Online Platform for Dictionary Publishers. In *Proceedings of the XIVth Euralex International Congress*, pages 389–401, Leeuwarden.

Lareau, F., Dras, M., Börschinger, B., and Turpin, M. (2012). Implementing Lexical Functions in XLE. In Butt, M. and King, T. H., editors, *Proceedings of the LFG12 Conference*, Stanford. CSLI.

Lux-Pogodalla, V. and Polguère, A. (2011). Construction of a French Lexical Network: Methodological Issues. In *Proceedings of the First InternationalWorkshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, pages 54–61, Ljubljana, Slovenia.

Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Wanner, L., editor, *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Language Companion Series*, pages 37–102. John Benjamins, Amsterdam/Philadelphia.

Mel'čuk, I. (2004a). Actants in semantics and syntax I: actants in semantics. *Linguistics*, 42(1):1–66.

Mel'čuk, I. (2004b). Actants in semantics and syntax II: actants in syntax. *Linguistics*, 42(2):247–291.

Mel'čuk, I. and Polguère, A. (2006). Dérivations sémantiques et collocations dans le DiCo/LAF. *Langue française*, 150:66–83.

Mel'čuk, I. and Polguère, A. (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.

Mel'čuk, I. and Žolkovskij, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna.

Mel'čuk, I. *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, Montreal.

Milićević, J. (2006). A Short Guide to the Meaning-Text Linguistic Theory. *Journal of Koralex*, 8:187–233.

Milićević, J. (2009). Schéma de régime : le pont entre le lexique et la grammaire. *Langages*, 176:94–116.

Polguère, A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In *Proceedings of EURALEX'2000*, pages 517–527, Stuttgart.

Polguère, A. (2007). Lessons from the *Lexique actif du français*. In Gerdes, K., Reuther, T., and Wanner, L., editors, *Meaning-Text Theory 2007. Proceedings of the Third International Conference on the Meaning Text Theory, Klagenfurt, May 20–24, 2007*, Wiener Slawistischer Almanach Sonderband 69, pages 397–405, München–Wien.

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

Polguère, A. (2012a). Lexicographie des dictionnaires virtuels. In Apresjan, Y., Boguslavsky, I., L'Homme, M.-C., Iomdin, L., Milićević, J., Polguère, A., and Wanner, L., editors, *Meanings, Texts, and Other Exciting Things. A Festschrift to Commemorate the 80th Anniversary of Professor Igor Alexandrovič Mel'čuk*, Studia Philologica, pages 509–523. Jazyki slavjanskoj kultury Publishers, Moscow.

Polguère, A. (forthcoming 2012b). Like a Lexicographer Weaving Her Lexical Network. In *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*. Summary of invited talk.

Polguère, A. and Sikora, D. (to appear ToAp). Modèle lexicographique de croissance du vocabulaire fondé sur un processus aléatoire, mais systématique. In Masseron, C., Garcia-Debanc, C., and Ronveaux, C., editors, *Enseigner le lexique. Pratiques sociales, objets à enseigner et pratiques d'enseignement*, volume 5. AiRDF.

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

Sagot, B. and Fišer, D. (2011). Extending wordnets by learning from multiple resources. In *Proceedings of LTC 2011*, Poznań.

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

Spohr, D. (2012). *Towards a Multifunctional Lexical Resource. Design and Implementation of a Graph-based Lexicon Model*. De Gruyter, Berlin/Boston.

Wolter, B. (2006). Lexical Network Structures and L2 Vocabulary Acquisition: The Role of L1 Lexical/Conceptual Knowledge. *Applied Linguistics*, 27(4):741–747.

Zock, M. and Schwab, D. (2011). Storage does not Guarantee Access: The Problem of Organizing and Accessing Words in a Speaker's Lexicon. *Journal of Cognitive Science*, 12:233–259.