Lessons from the *Lexique actif du français*

Alain Polguère

OLST—Département de linguistique et de traduction,
Université de Montréal
C.P. 6128, succ. Centre-ville,
Montréal (Québec) H3C 3J7 Canada
alain.polguere@umontreal.ca

Abstract

The *Lexique actif du français* (LAF) lexicographic project has just been completed. In its published form, the LAF contains the description of some 20,000 French semantic derivations and collocations. The theoretical and descriptive approach underlying the LAF is the Explanatory Combinatorial Lexicology, the lexicology branch of Meaning-Text linguistics. The main originality of this piece of work lies in the fact that the explanatory combinatorial formal approach has been interfaced, or popularized, in order to make the description accessible to non-linguists (mainly, language teachers and students). The aim of this paper is to offer a general presentation of the LAF, with special attention paid to how tensions generated by the dual need to rigorously describe lexical relations and to make accessible these descriptions to the layman have been handled. Two descriptive devices used in the LAF will be examined: the hierarchy of semantic labels and the metalinguistic formulas encoding lexical function relations.

Keywords

applied linguistics, Explanatory Combinatorial Lexicology, semantic derivation, collocation, semantic label, lexical function, lexical database, DiCo, LAF

1 Introduction

The *Lexique actif du français* (hereafter, LAF) is a lexicographic project undertaken some ten years ago by Igor Mel’čuk and myself; the project has just been completed. The LAF was a time-limited project, that targeted one specific lexicographic “product” (Mel’čuk & Polguère, 2007); it is however a sibling of an on-going project: the construction of the DiCo database of French semantic derivations and collocations. This latter is an open work, that should be indefinitely continued through various mutations. The main purpose of this paper is not so much to present the LAF/DiCo project itself, which has already been taken care of in several publications (Polguère, 2000a; Polguère, 2000b; Mel’čuk & Polguère, 2006). Rather, I wish to draw lessons from the LAF achievement, paying special attention to one specific topic, namely: the interaction between constraints imposed to modelling by formal linguistic approaches and constraints stemming from notional transfer in the context of applied linguistics. It has been com-
common practice among formal linguists, since the heyday of Artificial Intelligence, to believe that computer natural language processing (NLP) is the ultimate application of formal linguistic models, against which they should all be measured for an evaluation of their appropriateness. In spite of the fact that I personally originate from the NLP community, I strongly believe this to be a total myth. Rather, I believe that only language teaching/learning application of formal linguistic models is a proper environment for testing the validity of such models, specially in regards to the lexicon. It is of course not the place to demonstrate or even discuss this credo. But I hope to provide some information on how, in the specific case of the LAF, much insight has been gained on the nature of lexical entities and their relations, insight that can improve the completeness and accuracy of formalization. As some opinions expressed here may not necessarily be shared by the two makers of the LAF, I will use the first singular person I to denote myself and the plural we to denote both I. Mel’čuk and myself, when I have reasons to believe I express a common perspective. I trust Mel’čuk’s voice will be loud enough to be heard by everyone if I happen to distort his own thoughts.

The structure of the paper is as follows: section 2 introduces the LAF project, focussing on the characteristics of the final product which have not been already described in previous publications; section 3 focuses on non-definitional semantic characterization, namely on the LAF hierarchy of semantic labels; section 4 is devoted to the encoding of lexical function relations by means of paraphrasing formulas; finally, section 5 concludes by summing up the various lessons learnt and explains how our lexicographic work will proceed from now on.

2 The Lexique actif du français lexicographic project

2.1 Pedagogical and descriptive goals

The LAF is not an actual dictionary. It is before all a lexicology manual, based on a practical, data-oriented approach. The first part of the volume is an introduction to general concepts of Explanatory Combinatorial Lexicology (ECL) and to the content and structure of the LAF micro-dictionary proper, which makes up the volume’s second part. This hybrid manual can be related to Picoche (1993) for its theoretical and lexicographic dual structure. However, contrary to Picoche (1993), the LAF gives the lion’s share to the lexicographic description (Part 2), around which the whole structure of the book is organized. Our targeted readers are language teachers who need to find theoretical and pedagogical guidelines for teaching vocabulary, as well as French data on which they can base their teaching strategies. Of course, we believe many other readers may be interested in making use of the LAF, among which language students and lecturers in linguistics.

The English translation of the full LAF title is Active lexicon of French. Vocabulary learning based on 20,000 semantic derivations and collocations of French. As indicated by the subtitle, the LAF is relevant to the field of applied linguistics. In this respect, a whole chapter of Part 1 (the manual proper) is devoted to proposals for setting up teaching activities. Three types of lexicon-related activities are considered.

1. Lexical exploration: compilation and study of mini-lexicons based on semantic fields; work on polysemic structure of vocables.
2. Linguistic production: development of paraphrastic capabilities; thematic writing with the help of mini-lexicons.

3. Lexicographic activity: lexicographic definition of lexical units; writing of LAF-style descriptions using analogy with already-existing entries.

As authors, we see the LAF before all as an extremely painful but enriching exercise of theory-to-practice gap-bridging. There would be much to say about how we decided to select and introduce the various ECL core concepts to non-specialists in the manual part of the LAF. But I will focus here mainly on the dictionary part—Part 2. It contains the description of 385 mono- and polysemic vocables, for a total of 780 lexical units. As indicated in the title, the bulk of the description lies in the modeling of some 20,000 lexical links: semantic derivations and collocations controlled by the headwords. For the purpose of semantic characterization of lexical units, the LAF makes use of a hierarchy of 675 distinct semantic labels. The size and richness of this hierarchy is a good indicator of the semantic diversity of the LAF wordlist. Another good indicator of this diversity is the fact that we have had to use no less than 254 semantic fields to semantically classify the LAF’s 780 headwords.

2.2 The LAF macro- and microstructure

The LAF’s macrostructure is rather conventional, similar to that of standard dictionaries. It is a set of alphabetically-ordered vocable entries, each entry containing a series of articles, one for each lexical unit in the vocable. In order to counterbalance this non-semantic macrostructure, the LAF contains annexes that give semantic access to lexical units through semantic labels and semantic fields. The originality of the LAF truly lies in its microstructure. Being developed according to ECL principles, it is as far in terms of microstructure from standard dictionaries as Explanatory Combinatorial Dictionaries or ECDs (Mel’čuk et al., 1984, 1988, 1992, 1999) can be. Additionally, being a layman’s lexicographic model, it has a design that drastically distinguishes it from “kosher” ECL lexicographic models. This will be made clearer in sections 3 (on semantic labels) and 4 (on the encoding of lexical function links).

Figure 1 below gives the LAF entry of the monosemic French vocable \( \text{\textcopyright BON SENS} \) ‘common sense’; it serves as illustration of the prototypical structure of all LAF articles. This structure can be analyzed as follows.

- A semantic label (\textit{faculté de raisonner}) is the entry point into the article, immediately followed by the headword’s predicative structure described in a propositional formula.

- This latter also serves as a means of presenting the government pattern, with possible surface-syntactic realizations of each semantic actant, given between square brackets.

- Then comes a first block of lexical links, those pertaining to synonymy and quasi-synonymy—introduced by the ☞ symbol.

---

1 A semantic navigation tool (the DiCoPop, developed by Sébastien Cabot) is available on the LAF companion website: http://olst.ling.umontreal.ca/laf.
• It is followed by the enumeration of all other lexical function links, where popularization formulas (in bold) are used in place of actual formal encoding.

• The article ends with lexicographic examples—which can be immediately followed by pointers to phraseological lexical units built with the headword (not present in this case).

**BON SENS**, locution nominale, masc; pas de pl

**FACULTÉ DE RAISONNER**

Bon sens de la personne X [= de N, A_poss]

☞ sens commun; sagesse; intelligence; jugement; perspicacité

Ant. bêtise Quantité de B. S. dose [de ~] Petite quantité de B. S. grain, once, un peu [de ~]

Développé grand | antépos; beaucoup [de ~] Relativement élémentaire gros | antépos [X] qui a du B.S. de [-] [C’est quelqu’un de bon sens.]

Quantité de B. S. avoir, posséder [du ~] [X] manifester son B. S. faire preuve, faire montrer [de ~ en \(V_{pprét}\)] [Il a fait preuve de bon sens en refusant de s’embarquer dans cette aventure.], manifester, montrer [du ~ en \(V_{pprét}\)]

*On pouvait se fier à son jugement et à son bon sens. C’est avant tout un problème de bon sens.*

Figure 1: LAF entry for “BON SENS” (common sense) (Mel’čuk & Polguère, 2007:132)

Readers familiar with French ECDs will have noticed that the general presentation of a LAF article is more “textual” than formal descriptions found in other ECL-based dictionaries. The reason for proceeding this way is the necessity to limit the amount of printed pages of the dictionary. In theory, we would have much preferred to use separate lines for each lexical function link. Some articles, that feature a huge number of links, have been structured with “sections” in the lexical function zone of the entry, each section carrying a title. For instance, the entry for MAISON 1 (“house”) has the following nine sections: House dimensions, House characteristics that are links to its use, House appearance, House characteristics that are linked to its mode of construction or its position, Localization of something in a house or of the house itself, Construction of a house, Use of a house, Components of a house, Type of houses. Of course, choices we have made in this specific case can be discussed and refined. But we believe this type of structuring is of great help for navigating through lexicographic data and we would have gladly used it systematically in all entries. One of the approach that could be used in the future is to explicitly structure the lexical function zone of ECL descriptions, making apparent a system, or several competing systems, of classification of lexical function links (Jousse et al., To appear). This approach is however more relevant to on-line, rather than printed lexicographic descriptions (see section 5).

### 2.3 Lexicographic methodology: the DiCo filiation

Even in traditional lexicography, the published dictionary is seldom a direct product of the lexicographic activity. First comes a collection of lexicographic records which, in computational lexicography, takes the form of a structured database; the actual dictionary is derived from these records. So it goes for the LAF, which is nothing but a by-product of a richer computerized lexicographic model: the DiCo. Figure 2 below is the lexical function zone (QSyn and FL fields) of the DiCo record for “BON SENS”, whose LAF entry was given in Figure 1.
Lessons from the “Lexique actif du français”

{QSyn} _sens commun_; sagesse; intelligence; jugement; perspicacité

{QAnti} bêtise
/*Quantité de B. S.*/
{Sing} dose [de ~]
/*Petite quantité de B. S.*/
{AntiMagnSing} grain, once, un peu [de ~]
/*Développé*/
{Magn} grand | antépos; beaucoup [de ~]
{Relativement élémentaire} gros | antépos
/*[X] qui a du B. S.*/
{A1} de [-] [“C’est quelqu’un de bon sens.”] < plein [de ~]
/*[X] avoir du B. S.*/
{Oper1} avoir, posséder [du ~]
/*[X] manifester son B. S.*/
{Real1} _faire preuve_, _faire montre_ [de ~ en V-pprés] [“Il a montré du bon sens en refusant de s’embarquer dans cette aventure.”],
manifeste, montrer [du ~ en V-pprés]

Figure 2: Lexical function relations in the DiCo record for “BON SENS”

For lack of space, I will not give a detailed account of the correspondence between DiCo data and actual LAF entries. Let me simply emphasize the fact that all information contained in a LAF entry is already present in the corresponding DiCo record. The process of DiCo-to-LAF conversion is one of (re-)formatting and simplification. Simplification appears in the above example, where the reader can easily notice that the dual—formal and paraphrastic—encoding of lexical relations in the DiCo gives place to a unique non-formal encoding in the LAF.

The work on the DiCo started (very slowly) and went through the following five stages:

1. launching of the construction of the DiCo some 13 years ago;
2. after a significant number of vocables had been modeled and the DiCo format and methodology have been made more or less stable, launching of the parallel construction of the LAF—LAF entries being manually produced from corresponding DiCo records;
3. painful progression, where more DiCo records are built and translated into LAF format making sure LAF and DiCo transcriptions are kept in-sync—each time a modification is made in the DiCo, it has to be introduced in the LAF, and vice versa;
4. launching of an SQL version of the DiCo, obtained by compiling the original DiCo FileMaker records into a searchable datastructure, where each basic piece of information is clearly identified as content of a cell in an SQL table (Steinlin et al., 2005);
5. completion of the LAF project and launching of a project targeting the encoding of the lexical model in the form of a graph structure called lexical system (Polguère, 2006).

The process of designing a format for LAF entries was not an easy one, and can be seen as a long series of genetic mutations from monstrous half-formal-half-popularized descriptions, to smoother dictionary entries that follow the pattern exemplified in Figure 1. We went through many stages, submitting tentative entries to external evaluators, to colleagues and to students. Needless to say that these entries were often harshly criticized. The main problem was to find
the proper balance between what I. Mel’čuk calls “(scientific) truth”—that I personally see as raw scientific logic—and intelligibility for the layman. Even though painful, this long process of proposing and refining LAF formats was a very constructive one. It allowed us to identify a rather small subset of essential ECL notions that could not be dispensed with in a dictionary of semantic derivation and collocations. These notions are: headword’s core semantic content, headword’s actancial structure, headword’s government pattern, semantico-syntactic value of paradigmatic and syntagmatic links and government pattern of the headword’s collocates. All these prove to be required if one wants to provide a lexicographic description that is both necessary and sufficient (though incomplete) in the context of vocabulary teaching and learning. I will now proceed with considerations on how (i) semantic characterization and (ii) encoding of lexical links have been implemented in the LAF (sections 3 and 4, respectively).

3 Semantic characterization of lexical units

The DiCo/LAF notion of semantic label was put on the working table at the very beginning of the DiCo project as a means of offering a basic semantic characterization of lexical units. The task of giving a solid theoretical basis to the notion was done in collaboration with J. Milićević (Milićević, 1997). We were then able to develop a core hierarchy of semantic labels (HSL) for French, that has never stopped growing and evolving since then (Polguère, 2003a).

It should be emphasize that a semantic label is not a “deep concept.” It is only a pointer to the meaning of a given lexical unit of the language or of an expression (when no lexicalization is available). In fact, the semantic label of a given lexical unit is nothing but the central component of its definition, its genus. The process of labeling a headword is therefore very lexicographic in nature: it boils down to determining whether the label in question is the proper starting point for the analytical definition of the headword. Originally, I believed it would be possible to identify for each label a set of “objective” linguistic tests—compatibility/incompatibility with expressions and constructions—that would guide us in the process of semantic labelling. This proved to be a wild goose chase for a very simple reason: language quite naturally allows us to overwrite some semantic properties of lexical units by positioning them in competing semantic environments. For instance, one may consider that the English verb GROW

\[ \text{Baby Marc has grown a lot.} \]

denotes a process and is therefore non-punctual in nature. But nothing forbids us from giving it a punctual aspect, as shown in the following sentence.

(1) All of a sudden, at the age of 13, he grew 10 centimeters.

After much trial and debate, it became quite clear that, in order to be rigorous, the process of semantic labelling should rely entirely on the process of partial definition of headwords. Because the default link between semantic labels is hyperonymy, the set of 676 semantic labels used in LAF is formally a huge hierarchized graphs, from the vaguer to the more specific semantic labels. The whole task of finding proper semantic labels lead us to postulate that an HES that would be appropriate both for NLP and applied linguistics should possess the five following characteristics.

1. It should be inductively built, i.e. derived from the lexicographic process.

\[ \text{On the notion of analytical definition, see Polguère (2003b:150-158).} \]
2. It is language-dependent. The LAF’s HSL is therefore intrinsically linked to French and no claim is made about its universality.

3. It is relatively flat, as hyperonymic links develop breadth-first. In spite of the high number of labels we used, it is very seldom the case that any given semantic label in the HSL possesses more than five ancestors until the root label is reached (qqch. ‘something’).

4. It is not a perfect tree as some labels may have more than one mother in the hierarchy. For instance, names of professions are all labelled individu qui pratique un métier ‘individual who has a profession’. This label has two mothers—individu ‘individual’ and métier ‘profession’—as any lexical unit that can be labelled with it can also be used to denote either an individual (He killed his doctor) or a profession (Doctor is a very challenging profession).

5. It incorporates non-hyperonymic links. For instance, basic nominal semantic labels govern in the hierarchy labels that are verbal, adjectival or adverbial semantic derivations.

This latter characteristic makes the hierarchization of lexical units in the LAF/DiCo very different from that offered by lexical models of the WordNet family. Firstly, unlike WordNet, the LAF/DiCo model is only implicitly hierarchized. Hierarchization of lexical units by means of the HSL is a projection of a given perspective on the described vocabulary. It is not a hard-coded structuring in our model (Polguère, 2006). Secondly, WordNet is based on a strict partition of the lexicon following major parts of speech. In this respect, the WordNet hierarchy is in fact the conjunction of four separate hierarchies (for nouns, verbs, adjectives and adverbs). Because we think that the proper classification and hierarchization of lexical units is before all semantic, rather than grammatical, we found it essential to integrate labels of all parts of speech within one single HSL. As a result, we are able to project onto the described vocabulary a hierarchy that unifies actual lexical meanings, disregarding grammatical constraints on how these meanings should be expressed.

To conclude on semantic labels, let us emphasize the fact even though they prove to be a very powerful tool for both theoretical and applied lexicology, they are not a full substitute for actual definitions. These latter will always be needed in studies in lexical semantics based on ECL (Barque & Nasr, 2005; Iordanskaja & Polguère, 2005) or in applied linguistics.

4 Metalinguistic encoding of lexical function relations

An advanced knowledge of core semantic and syntactic notions is required in order to understand standard lexical function formulas. For instance, in order to interpret the following lexical functions found in the DiCo entry for ‘BON SENS’, one has to be able to “visualize” the semantic graph and the syntactic dependency tree associated with the expressions avoir/posséder du bon sens vs. faire preuve/FAIRE MONTE DE BON SENS (Polguère, 2003:139-142).

(2) {Oper1} avoir, posséder [du ~]
{Real1} _faire preuve_, _faire montrer_ [de ~ en V-pprés]

Because we could not expect such advanced knowledge of linguistics from our targeted readers, it was decided to use in the LAF only so-called formules de vulgarisation ‘popularization formulas.’ The main idea behind the use of such formulas is to treat each standard lexical func-
tion link as if it were non-standard, and account for it by means of the same type of defining expression that is used in ECDS for non-standard links. Additionally, because standard links are by definition recurrent, we expected to be able to normalize popularization formulas for all such standard links. It quickly became apparent that there was a strong conceptual connection between the choice of proper popularization formulas for standard links and the semantic labels of the arguments of lexical functions. For instance, most nouns of feelings could get the same popularization formulas for their \texttt{Magn}, \texttt{Oper}_1, etc.

At a more general level, the work on popularization formulas gave us much insight on the potential role of paraphrase in the modeling of lexical function links (Polguère, 2004). Because standard lexical functions play a central role in paraphrasing (Miličević, 2007), their translation in controlled natural language is itself a powerful tool for reasoning on lexical links from the perspective of paraphrase. In the manual part of the LAF (page 59), for instance, we use data (3) taken from the entry of FÉLICITATIONS \texttt{a} \texttt{congratulations} \texttt{b} \texttt{Il a reçu les félicitations du jury.} to show how lexical relations that may appear to be very different at first glance for the layman are in fact involved in the expression of sentences of similar content, in (4).

(3) \texttt{Verbe} [= \texttt{V}_0] féliciter [N\texttt{V} \texttt{Prép}_\texttt{a propos} N\texttt{Z}]
\texttt{Ce que X dit comme F.} [= \texttt{Proposition}] « Félicitations ! »
\texttt{Lettre par laquelle X communique des F.} [= \texttt{s}_\texttt{med}] lettre [de ~]
\texttt{[X] communiquer des F. à Y} [= \texttt{Oper}_1] adresser, exprimer [ART \texttt{à N\texttt{V} \texttt{Prép}_\texttt{à propos} N\texttt{Z}]}\texttt{][Y] être vist par les F. de X} [= \texttt{Oper}_2] recevoir [ART \texttt{de N\texttt{X} \texttt{Prép}_\texttt{à propos} N\texttt{Z}]}\texttt{][Z] être la raison de F.} [= \texttt{Oper}_3] valoir [ART \texttt{à N\texttt{V}]}\texttt{]

In order to make more visible paraphrasing relations implied by the LAF data, I have inserted the corresponding lexical function names in the LAF description in (3).

(4) \texttt{Alain a félicité Marc pour son exploit.}
« Félicitations pour ton exploit ! », a dit Alain à Marc.
\texttt{Alain a envoyé une lettre de félicitations à Marc à propos de son exploit.}
\texttt{Alain a adressé, exprimé des félicitations à Marc à propos de son exploit.}
\texttt{Marc a reçu des félicitations d'Alain à propos de son exploit.}
\texttt{Son exploit a valu à Marc les félicitations d'Alain.}

Interestingly, the use of popularization formulas led us to introduce so-called locally standard lexical functions in the DiCo (Polguère, 2007) in order to account for the recurrent patterns of non-standard lexical links. For instance, the locally standard lexical function \texttt{Proposition} in (3) was introduced after A.-L. Jousse noticed that \texttt{Ce que X dit...} ‘What X says…?’ happened to be quite common as popularization formula in the DiCo. There would be much to say about all avenues that the use of popularization formulas opens in both theoretical and applied ECL but, for lack of space, I have to jump directly to my conclusions.

5 Conclusions

The process of popularization and “simplifying” standard ECL models has led us to develop descriptive tools that have practical \texttt{and} theoretical advantages of their own. In some respect, semantic labelling proves to be a more powerful way of semantically characterizing lexical units than actual analytical definitions. Of course, as mentioned earlier, it leaves many things
Lessons from the “Lexique actif du français”

aside; but a good semantic labelling together with a specification of the predicative structure of the lexical unit summarizes the semantic essence of the headword and is a powerful tool for any operation that involves generalization. In some sense, the semantic description used in the LAF may have applications similar to that of the FrameNet approach (Baker et al., 2003). Similarly, the paraphrasing of lexical function relations bridges the gap between standard and non-standard lexical functions. It allows for new kinds of generalizations and is an essential tool in the exploration of the nature of lexical function standardness (Polguère, 2007).

So what comes next? Firstly, we will make a follow-up of the LAF project by developing its companion website (see footnote 1). Secondly, the process of converting the DiCo database into a lexical system (Polguère, 2006) has just started. I hope that this will allow us to explore and put in place more sophisticated lexicographic approaches than what has been done until now within the ECL framework. Specific avenues that will be explored are: “intelligent” editing of lexical entries (with consistency checks), automatic generation of dictionary entries on the fly directly from the database and according to specific user needs3 and, finally, automatic generation of draft entries from already existing ones—using an analogy-based algorithm.

Acknowledgments

Work on the DiCo and LAF projects has been partly supported by research grants from the FQRSC (Quebec) and SSHRC (Canada) agencies.

Bibliography


3 M.-C. L’Homme, in collaboration with A. Benoît et G. Lapalme, is already working on the automatic generation of an XML-PDF terminology dictionary directly from lexicographic tables (L’Homme, In preparation).
Alain Polguère


