A practical guide to writing *DiCo* entries

François Lareau
francois.lareau@umontreal.ca
OLST — Département de linguistique et traduction, Université de Montréal
www.olst.umontreal.ca

Introduction

This paper is intended for people who are already somewhat familiar with the *DiCo* project (its formalism and the basic notions on which it is based). It does not offer a detailed presentation of this lexical database nor does it discuss theoretical issues in lexicography. For such discussions, the reader should consult (Polguère, 2000a), (Polguère 2000b), (Mel'čuk, Clas and Polguère, 1995 — hereafter, *ILEC*), and (Mel'čuk et al., 1984, 1988, 1992, 1999 — hereafter, *DECFC)*. The purpose of this paper is rather to offer a practical, step-by-step guide to writing *DiCo* entries, as a complement to *ILEC*.

The steps listed below are roughly those followed at the OLST[1]. The description of vocables is done in two phases: first, give an approximate description (this task being carried out mostly by students), and then refine the descriptions (this being done by more experienced lexicographers).

Step 1: Choosing a lexical field

Though it is still common practice in modern lexicography to proceed in alphabetical order, it offers no advantage at all. On the contrary, it often leads to confusion and mistakes. A better approach is to proceed by lexical fields. A lexical field is a set of vocables whose basic lexical units[2] are semantically close (see the example below). The choice of the lexical field to study is up to the lexicographer.

The main reason for doing so is that lexical units which belong to the same lexical field often share a lot of common information. They are very likely to have similar semantic formulas, similar government patterns, and common collocations. Describing them together, in parallel, will help the lexicographer a lot. It helps ensuring that no similar information from a lexical unit has been left out in another, and that the differences between lexical units appear clearly. It also helps keeping lexical descriptions homogeneous. For a more detailed argumentation, see *ILEC*, pp. 181-184.

For instance, the lexicographer might want to describe the lexical unit BATAILLE ‹battle›, as in *Les troupes du général Polguère ont gagné la bataille*. Then, he will have to consider all vocables of which the basic lexical unit is semantically related to ‹battle›: ARMÉE ‹army›, BAGARRE ‹fight›, SE BAGARRER ‹to fight›, VICTOIRE ‹victory›, etc. The set of all these vocables is the lexical field of ‹battle›. For each of these vocables, the following steps are followed.

Step 2: Dividing the vocables into lexical units

When describing a vocable, the lexicographer's first task will be to separate the vocable into lexical units. For this purpose, corpora are very useful, if not necessary: Alone, the lexicographer's knowledge of the language considered will probably not suffice, for some facts might simply not come to his mind spontaneously. For a rigorous and detailed presentation of the principles to follow when dividing a vocable in lexical units, see *ILEC*, pp. 56-69 and pp. 155-171.

This step might lead to distinguishing homonymic vocables. If what seems to be two senses of a vocable share no significant semantic component, that means they are **not** lexical units of the same vocable, but rather belong to two distinct, homonymic vocables.

1. Observatoire de linguistique Sens-Texte, where the *DiCo* is being developed.
2. See *ILEC*, p. 159, for a definition of the concept of basic lexical unit.

For instance, when describing the English word *character*, one may find the following examples in a corpus:

a) *As we flew North, the character of the country changed drastically.*
b) *This guy has a very good character.*
c) *I can't find the character "ñ" on my keyboard.*

While ***a)*** and ***b)*** are semantically related, ***c)*** has nothing to do with these meanings. The lexicographer will therefore postulate two homonymic vocables, CHARACTER[1] and CHARACTER[2] (the order is not important), one of which being polysemic (regrouping the senses ***a)*** and ***b)***).

When a vocable is polysemic, a separate database record is created for each lexical unit, *i.e.* each acception of the vocable. Lexical units are numbered according to their semantic "distance" (this numbering might be at first approximate; it will be refined later). Also, a separate record is created and numbered "0". This "zero-record" will be used to store the information that is shared by all lexical units of the vocable (see below, step 7). For the moment, one could at least indicate in this "zero-record" the part-of-speech of the vocable. For each lexical unit of a vocable, the following steps are applied.

Step 4: Finding synonyms and giving examples

As soon as the lexicographer distinguishes a new lexical unit, it is good practice to write down a few synonyms and examples (from the corpus) in the appropriate fields of the record. This will avoid eventual confusion when coming back to the record for further work (especially if this is done by another lexicographer). It helps remembering what meaning the lexicographer had in mind when creating this record and prevents hair loss.

Step 5: Establishing the semantic structure

Prior to establishing the government pattern and collocations of a lexical unit, one needs to know what is its semantic structure. For this reason, the "cs" field (*caractéristiques sémantiques* — semantic characteristics) has to be filled out first. It is not necessary to determine right now what the semantic labels of the lexical unit and its semantic actants are. What is important for now is to establish 1) wether the lexical unit under analysis is a predicate or not, and 2) if yes, how many semantic actants does it have? This first estimation is only an hypothesis on which to base the rest of the work, and is subject to changes when the examples and collocations will be examined. The reader might want to have a look at *ILEC*, pp. 75-78, for a clear presentation of the notion of semantic actant.

Step 6: Drafting the government pattern

The government pattern of a lexical unit indicates how to map the semantic actants [=SemAct] to deep-syntactic actants [=DSyntAct] as well as the possible surface realizations for these actants. This information is encoded as follows (the basic pattern given here is that of the French noun ASSASSINAT ‹assassination›):

```
X = I = par N
Y = II = de N, A-poss
```

The above table (*DiCo* tables are equivalent to the ones found in the *DECFC*, see *ILEC*, pp. 117-123 and p. 221) says that the first SemAct is mapped to the first DSyntAct, and is further realized as a noun phrase introduced by the preposition PAR (equivalent[1] to the English BY), while the second SemAct should be the second DSyntAct and can be realized as either a noun phrase introduced by DE (equivalent to the English OF) or an *adjectif possessif* (A-poss), such as MON ‹my›, SON ‹his/her›, LEUR ‹their›…

In languages where the nouns are marked with a case, the cases assigned by a lexical unit

1. Note that saying they are equivalent does not mean they have the same ***meaning***. These prepositions, whose choice is controled by the keyword, have no meaning, or they are emptied from their original meaning; they play a strictly syntactic role.

to its actants have to be indicated in the government pattern. For example, a verb for which the first SemAct has to be realized as a noun in the nominative while its second SemAct is to be in the dative, would have the following government pattern:

```
X = I = N-nom
Y = II = N-dat
```

In most cases, writing a government pattern is quite simple. The mapping between SemActs and DSyntActs is usually trivial (X=I, Y=II, Z=III...), and the surface realizations are easily available to the lexicographer. Moreover, it seems that there are not so many different patterns in a given language. However, there are some more complex cases…

There might be more than one possible mapping between the SemActs and DSyntActs of a lexical unit. In that case, we have to give more than one table. For instance, in the *DiCo* record for BATAILLE$_{I.1}$, we have:

```
Mod. 1
X+Y = I+II = entre N et N, A-poss

Mod. 2
X = I = de N, A-poss
Y = II = contre N
```

In the case of Mod. 1[1], the SemActs 1 and 2 are realized together as only one DSyntAct labelled "I+II" (*La bataille entre* [*les Anglais et les Français*]$_{I+II}$ or [*Leur*]$_{I+II}$ *bataille*), while in Mod. 2, they are both expressed as separate DSyntActs (*La bataille* [*des Anglais*]$_{I}$ *contre* [*les Français*]$_{II}$ or [*Leur*]$_{I}$ *bataille contre* [*les Français*]$_{II}$).

At this point, we have some kind of an X-ray picture of the lexical unit's skeleton. The record already contains enough information to be roughly usable by a computer (it is far from being complete, but at least the necessary information on how to handle it is there), and most of all, the lexicographer now has more solid ground on which to proceed.

Step 7: Building a list of semantic derivations and collocations

The very important concept of lexical function [=LF] is extensively discussed in *ILEC*, pp. 125-152.

Describing the semantic derivations and collocations of a lexical unit takes a lot of time. Each lexicographer might have his own strategy, but what seems to be a very efficient way is as follows:

1) List all semantic derivations and collocations, along with their government pattern. To find collocations of a lexical unit, one can use dictionaries such as (Lacroix, 1947) or (Rouaix, 1997), for French, or (Benson, Benson and Ilson, 1986), for English, which give exactly that: a list of collocations for each lexical unit. Corpora are also a very good source of information, for they give not only the collocations themselves, but also their government pattern (since they are used in context).

2) Group the semantic derivations and collocations according to their meaning and syntactic behavior. This amounts to putting together expressions believed to be values of the same LF.

3) Try to figure out which LF the groups of expressions are values of. Very often, it will not be an easy task. The lexicographer should not try to get a complete and perfect description of LFs at once. It is much easier to just put a question mark when not sure, and later come back on the problem, or leave it to a more experienced lexicographer. Students working at the OLST will typically hand in something like (again, for BATAILLE$_{I.1}$):

```
{?} coup
{?} batailleur
{?} agressif
{Magn} grosse, rangée
{?} de rue
{Loc-in} au cours de, durant,
         pendant [ART ~]; dans
         [ART ~]
```

1. This *Mod.* stands for *modification* and should simply be understood as one possible pattern. One could write *Pattern 1*, *Pattern 2*, … instead of *Mod. 1*, *Mod. 2*, …

```
{?} en pleine [~]
{Oper1} participer [à ART ~]
```

Step 8: Extracting information shared by all lexical units of a vocable

Once a (polysemic) vocable's lexical units have been described, the lexicographer will sometimes find that some information is shared by all of its lexical units. This information should be removed from the lexical units' entries to be put in the "zero-record" that was created at step 2. Such information will typically be elements of the syntactics (especially, the part-of-speech is usually the same for all lexical units of a vocable) and common standard lexical functions. It is also possible that all lexical units of a vocable will have the same government pattern). Note that it is absolutely impossible to have anything in the semantic or phraseology fields of the "zero-record", since these are specific to each lexical unit.

Step 9: Completing the semantic structure

Specifying the semantic label (more or less, the genus, *i.e.* the semantic core) of a lexical unit and those of its semantic actants is a bit similar to writing a definition. The lexicographer will need for this task a hierarchized list of semantic labels for the given language (this list being in perpetual evolution). There is currently such a hierarchy for French, but not for other languages (though it must be possible to use a modified version of an existing semantic hierarchy). For much more about semantic labels, see (Milićević, 1997).

Step 10:Refining

When all this is done, we have a good draft. The entries are then passed over to colleagues, who will go through the steps 2 to 9 over and over again, always making changes, adding or correcting information, etc. When "revisiting" a database record, the lexicographer will consider it in parallel with other entries to make sure that descriptions are consistent.

Acknowledgements

Many thanks to Alain Polguère, whose numerous corrections and comments on a previous version of this paper were very helpful. I am of course the only one to blame for remaining errors and inconsistencies.

References

Benson, M., E. Benson and R. Ilson. 1986. *The BBI combinatory dictionary of English: A guide to word combinations*. John Benjamins, Amsterdam/ Philadelphia.

Lacroix, U. 1947. *Dictionnaire des mots et des idées: Les idées par les mots*. Nathan, Paris.

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

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

Milićević, J. 1997. *Étiquettes sémantiques dans un dictionnaire formalisé du type Dictionnaire Explicatif et Combinatoire*. Mémoire de maîtrise, Département de linguistique et de traduction, Université de Montréal.

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

Polguère, A. 2000b. *Une base de données lexicales du français et ses applications possibles en didactique*. In *Revue de Linguistique et de Didactique des Langues (LIDIL)*, 21: 75-97.

Rouaix, P. 1997. *Trouver le mot juste - Dictionnaire des idées suggérées par les mots*. Le Livre de Poche no. 7939, Armand Colin, Paris.

Symbols and abbreviations used

‹ ... ›	= meaning
SMALL CAPS	= lexical unit / vocable
Italics	= example
`Courier`	= extract from the *DiCo*
LF	= lexical function
SemAct	= semantic actant
DSyntAct	= deep-syntactic actant