Collecting, Encoding and Organizing Collocates in a Terminological Database

Marie-Claude L'Homme
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal

This paper will present a methodology for collecting collocates of terms in a specialized domain, and for encoding and organizing them in a terminological database. The methodology was developed and implemented in a French dictionary on computing and the Internet (DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet) which contains approximately 2000 articles (each articles corresponds to a specific sense).

First, we will describe the basic principles upon which our terminological database relies, namely a lexico-semantic framework largely based on Explanatory Combinatorial Lexicology, ECL (Mel'čuk et al. 1985, 1984-1999). We will also give a brief description of our general methodology and computer tools used by terminologists working on the project: 1) compilation of the corpus; 2) automatic and manual selection of terms; 3) sense distinction; 4) definition of the actantial structure of terms; 5) listing of terms sharing a paradigmatic or syntagmatic relationship with the headword.

We will then focus on the work regarding syntagmatic relationships, namely the collection, encoding and organization of collocates in our terminological database. We will review some similar endeavors (specialized dictionaries containing information on collocates of terms, such as Cohen 1986; Binon et al. 2000; Meynard 2000) and examine how data regarding collocates is organized. We will also present the choices we made in our own database and its specificities in relation with previous work.

Collection: Collocates are first collected manually in a 1 million word corpus using the headword to obtain concordances. For example, for a noun term, terminologists search the corpus to extract adjectives, nouns and verbs which combine frequently with it and which express a relevant relationship (relevance is defined according to the lexical functions, LF, framework). A search in the Web can also be performed to make sure collocates have not been forgotten. Part of the list of collocates for the term *Internet* is given below :

> *Internet*: ~ *sans fil* (wireless Internet), ~ *haute vitesse* (high speed Internet), *se connecter à l'~* (to connect to the Internet), *naviguer dans l'~* (to surf the Internet); *rechercher qqch. dans l'~* (to search something on the Internet).

Encoding: Collocates are then listed under each word and the relationship shared between the collocate and the key word is encoded using two different methods. First, a formal encoding is performed using lexical functions (Mel'čuk et al. 1985, 1984-1999). LFs help terminologists systematize encodings throughout the database. For exemple, the LF Real_i encodes verbal collocates which convey the meaning of "typical use made of something"). Even, if key words and collocates differ, the formal encoding remains the same, as illustrated below:

Real$_1$ (Internet): *naviguer dans l'~* (to surf the Internet)
Real$_1$ (formulaire): *remplir un* ~ (to fill a data entry form)
Real$_1$ (fenêtre): *ouvrir une* ~ (to open a window)
Real$_1$ (touche): *appuyer, enfoncer une* ~ (to press, hit a key)

Then, an explanation. which refers to the the actancial structure of terms given at the beginning of the article, is given for each collocate. The encoding with LFs are used solely by terminologists working on the database; on the other hand, the explanation can be consulted by users of the database. An example is given below with *rechercher qqch. dans l'~* (to search something on the Internet), which shows are actants are used to write this explanation:

Internet : ~ utilisé par AGENT{internaute 1} pour intervenir sur PATIENT{ressource 2; site 1} avec INSTRUMENT{navigateur 1}

L'agent utilise un « mot clé » pour intervenir sur le patient : *rechercher un patient dans l'Internet*

Work has been carried out in order to convert this explanation into a more user-friendly one. Instead of using labels for semantic roles (agent, patient, etc.), we are planning on using the typical terms which instanciate this actant (e.g., *l'internaute utilise l'Internet pour intervenir sur la ressource ou le site*: the web user uses the Internet to act on the site or the resource).

Organization: Collocates are organizing within the article using the following criteria. First, collocates which specify a "type of key word" are listed. These are collocates which take the form of adjectives, nouns or preposition phrases (e.g., *Internet*: *~ haute vitesse, haut débit, sans fil*) (for noun terms). Then, verbal collocates are listed (together with nominalizations and adjectivations of verbs). An attempt is made to order verbs according to the sequence of use of the key word and to the complexity of the relationship between the collocate and the term. Part of this ordering is given for Internet below:

Se connecter à l'~ (to connect to the Internet)
Naviguer dans l'~ (to surf the Internet)
Se déconnecter de l'~ (lit. To disconnect from the Internet)
Chercher qqch. dans l'~ (to search something on the Internet)
Publier un site dans l'~ (lit. To publish a site on the Internet; to put a site on line)

We will conclude with a few remarks on the challenges this encoding presents and on future work.