# Advanced Encoding for Multilingual Access in a Terminological Data Base – A Matter of Balance

**Marie-Claude L'Homme[1], Patrick Leroyer[2] and Benoit Robichaud[1]**

[1] Observatoire de linguistique Sens-Texte (OLST), Université de Montréal
C.P. 6128, succ. Centre-ville, Montréal (Québec), H3C 3J7
mc.lhomme@umontreal.ca, benoit.robichaud@umontreal.ca
[2] Centlex, Aarhus School of Business, Aarhus University
Fuglesangs Allé 4, DK-8210 Aarhus V
pl@asb.dk

This paper describes new functionalities implemented in a terminological database (TDB) in order to allow efficient editing of and access to multilingual data. The functionalities are original in the sense that they allow users of the database to retrieve the equivalents not only of headwords, but also of semantically related terms (especially collocations) that appear within the articles. The methods we developed are based on a formal encoding of lexical relationships, namely lexical functions (LFs). Since they are language-independent and are designed to capture semantic distinctions, links between equivalents can be established automatically.

Keywords: terminological database, semantic relationships, collocations, access methods

## 1. Introduction

Terminological databases (= TDB) for natural language applications are information retrieval systems that (should) have been developed with the objective of truly helping their envisaged users to solve problems in foreseen communication situations: reading, translating or producing specialized L1-L2 texts. In commercial TDB’s designed by companies for the market, user-need adaptation is an essential functional parameter. It is also the case of TDB’s developed as research tools and based on an advanced encoding of the language data, the potential user group being then restricted to the research community. In this regard, it might be argued that extending user access to multilingual data could be regarded as some kind of compensation work. Nothing could be further from the truth. Developing a multilingual TDB reveals the potentials of advanced data encoding:

1. From an editing perspective – providing translation without translating the data directly (this is especially important in the case of the database used in

this work, since it is under construction and the descriptions have not reached the same degree of completeness in all languages);

2. From user's perspective – better access to the data

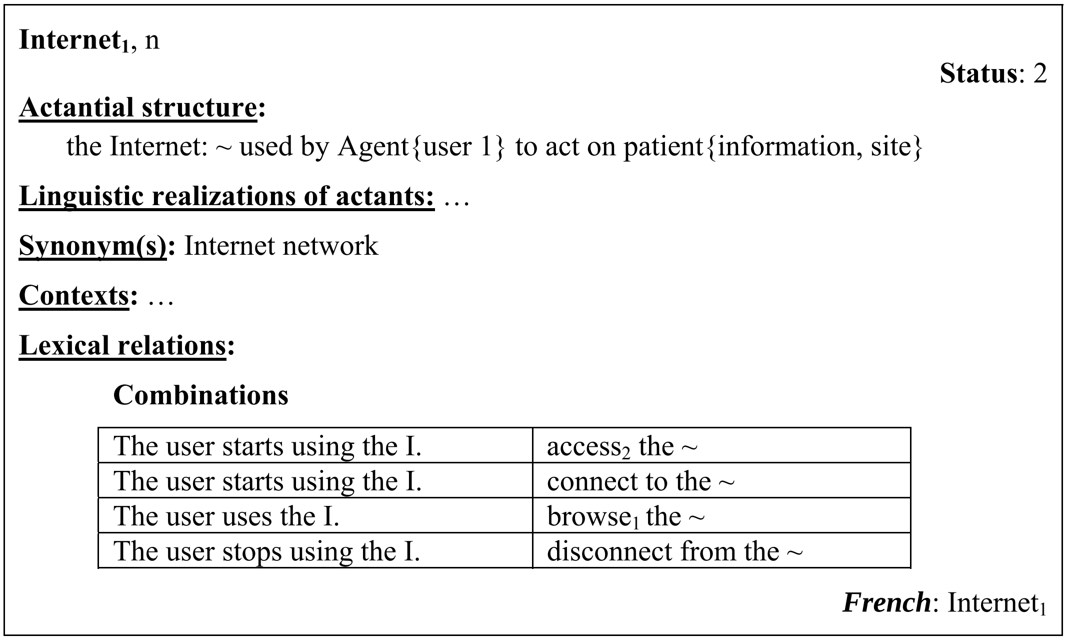
In this contribution, we present the new translation functionalities that are being developed for the DiCoInfo, a multilingual TDB based on the representation of the language of computer science and the Internet. We show that the two perspectives are intertwined and can be perfectly balanced.

The article is organized as follows. Section 2 gives a short description of the data encoded in the DiCoInfo and how it can be viewed in the website. In section 3, we argue that different translation functions and the needs they raise can be fulfilled by adding functionalities to the DiCoInfo. Section 4 describes how these functionalities are implemented. Finally, a few concluding remarks and directions for future work are given in Section 5.

## 2. The DiCoInfo

The DiCoInfo (*Dictionnaire fondamental de l'informatique et de l'Internet*) is an online TBD that contains terms in the fields of computing and the Internet (http://olst.ling.umontreal.ca/dicoinfo/). It provides rich linguistic information on terms and thus differs from most terminological databases which are usually concept-based. The main data categories in the DiCoInfo are the following (an example is given in Figure 1 below; the complete entry appears in the Appendix):

- Headword
- Grammatical information
- Actantial structure (also called argument structure)
- Linguistic realizations of actants
- Contexts
- Lexical relations (lists of paradigmatically and syntagmatically related units)
- Equivalents



**Internet$_1$**, n

**Status**: 2

**Actantial structure:**

the Internet: ~ used by Agent{user 1} to act on patient{information, site}

**Linguistic realizations of actants:** …

**Synonym(s):** Internet network

**Contexts:** …

**Lexical relations:**

**Combinations**

| | |
|---|---|
| The user starts using the I. | access$_2$ the ~ |
| The user starts using the I. | connect to the ~ |
| The user uses the I. | browse$_1$ the ~ |
| The user stops using the I. | disconnect from the ~ |

***French***: Internet$_1$

Figure 1: Part of the entry for *Internet$_1$*

The compiling of the DiCoInfo is based on a methodology combining computer tools and resources (a 1,000,000 word specialized corpus, a term extractor, a concordancer that takes into account POS tagging) along with analyses resorting to lexical semantics criteria (L'Homme 2008). The editing of articles is performed in the XML editor oXygen.[1] The main lexical framework on which the analyses and encoding are based is Explanatory Combinatorial Lexicology, ECL (Mel'čuk et al. 1984-1999, 1995).

Originally, the database was designed in a semi-formal way, but recently we have tried to convert some of its data categories and access paths in order to present them in a more user-friendly interface. The first proposal in that respect seems to yield promising results (L'Homme and Leroyer 2009).

Currently, the DiCoInfo contains over 1,000 articles for French terms (an article describes one specific sense). We started adding English terms to the database and it now contains a little over 200 articles for English dealing mostly with nouns, verbs and adjectives (ex. *attack, backup, character, configure, downloadable*). We must

[1] The schema for the article was designed in collaboration with Guy Lapalme from the Computer Science Department of the University of Montreal.

underline that the descriptive work is carried out on each language separately. However, when adding an English term, we link the French equivalent to the English term and vice versa.

Adding data in other languages than French[2] made us realize that it would be extremely profitable not only to link the headwords but also the various lexical relationships that appear within the articles (for instance, between combinations, cf. Figure 1). Hence, we developed a method for handling the links between records in different languages automatically without having to provide specific translations for each related unit. In addition, we wanted to achieve this without burdening users with very complex access paths or difficult to read formalisms.

## 3. Adapting data and access to user-oriented translation tasks

All the components of the DiCoInfo are presently in the process of being adapted and updated in order to optimize the efficiency of the translation functions.

For the L1 and the L2 reception phases of translation, and as far as headwords are concerned, the DiCoInfo provides a fairly comprehensive coverage of the domain. Besides headwords, it provides definitions and indexed access to lists of semantically related items.

For the L2 production phase, the presentation of grammatical data (actantial structures and linguistic forms of actants) is particularly valuable in combination with the semantics of lexical relations. Contexts provide useful pragmatic and stylistic information for the insertion of terms in the professional discourse.

For the L1-L2 translation phase itself, equivalents to the headwords are being systematically provided together with new access paths allowing direct access to equivalent collocations and other lexical relationships.

The interface (search page and display pages) can be customized to a certain extent by the user according to her/his translation needs, but access will be further improved by a clearly translation dependent search and display mode. Furthermore,

[2] A Spanish component is also planned to be added in the near future.

it is envisaged to include an external search path allowing users to enter expressions and browse the Internet.

## 4. Advanced encoding for multilingual access

As was said in Section 2, the main part of the article in the DiCoInfo is devoted to the description of lexical relationships the headword has with other terms of the dictionary (or with other lexical units that do not appear in the nomenclature). The relationships include paradigmatic relations (such as near synonymy, antonymy, hyperonymy) and syntagmatic relations (i.e., collocations).

It soon appeared interesting to link equivalents not only at the level of headwords, but also to display all the potential equivalences that could be established between entries in different languages at the level of lexical relations. In the DiCoInfo, since lexically related terms are all encoded using a system that describes their semantics, the equivalence relationships can be established without having to translate each of them. This section will show how we manage to extract the equivalents of lexically related terms based on the semantic encoding. We will first say a few words on the system used to represent lexically related terms, i.e. lexical functions. Then we will proceed to explain how this system is exploited to extract relationships automatically.

### 4.1 Encoding of lexical relations

In accordance with ECL, lexical relations are described using the systems of lexical functions, LFs (Mel'čuk et al. 1984-1999, 1995). LFs are language-independent and can express:

- A semantic relation (in the case of many paradigmatic LFs): e.g., **QSyn** (for near synonymy) is used to represent the relation between *run* and *execute*.
- A basic and general meaning: e.g., **Able** is used to represent the meaning "that can be verbed" between *compile* and *compilable*; **Gener** is used to represent the hyperonymic relationships between *Internet* and *network*.

- The syntactic function of the keyword (for collocations): e.g., **Real** is used when the keyword is first complement (*browse the <u>Internet</u>*); **Labreal** is used when the keyword is second complement (*enter data on a <u>keyboard</u>*).
- The actant involved in the semantic relationship: e.g., the subject of *browse* in *browse the Internet* is the first actant of Internet (cf. Figure 1); the LF will refer to this argument using a numbering system, i.e. **Real**$_1$.

The web version of the DiCoInfo does not display LFs in most entries, using rather a system of paraphrases that explain LFs in a more user-friendly way (Polguère 2003). **Real**$_1$, for instance, is often paraphrased as “The user (typical agent) uses the keyword”. However, terminologists, when encoding lexical relationships will do so using the system of LFs. As we will see later on, this system is useful to link equivalent lexical relations, and especially collocations.

### 4.2 Linking lexically related terms automatically

Because of their rich linguistic expressiveness and since they are language-independent, LFs are especially useful to link equivalent collocations. The computational strategy to do so is quite straightforward and twofold. First, according to the search options selected by users in the web interface, the TDB is queried to retrieve the set of all articles containing:

1. Entries that fulfill the user’s query (that is, entries that contain the searched expression in at least one lexical relationship description).
2. Entries of the corresponding equivalents.
3. Entries that are pointed out by the LFs.

Secondly, when displaying the search results, according to each lexical relation that has been found, Boolean and set constraints are checked to verify if a corresponding lexical relationship exists in the entry of the equivalent. We have to further distinguish the following:

1. In the case of syntagmatic relations, only the presence of the same LF describing a lexical relation is necessary and sufficient to decide that it is the corresponding equivalent collocation. This strategy is illustrated in Figure 2. These results are obtained when a user searches for the verb *browse*. The system extracts English collocations in which *browse* appears, but can also ascertain the French corresponding collocations based on the fact that they are encoded with the same LF.

**Combinations:**

**browser$_1$** : *browse$_1$* ... with a ~ (The internaut uses a b. to act on the internet)
(↔ fr: **navigateur$_1$** : *naviguer$_1$* dans ... avec un ~ (L'internaute utilise un n. pour intervenir dans Internet))
*Both collocations are encoded with the* **Labreal$_{12}$** *LF.*

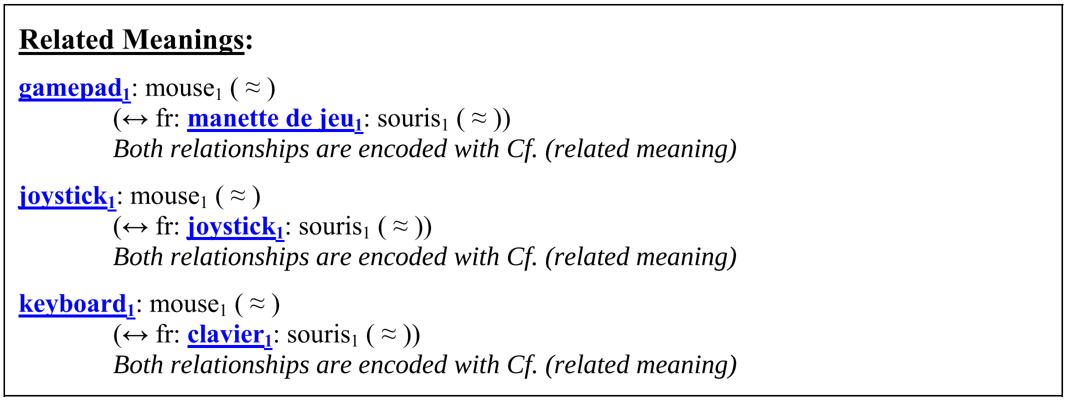
**Internet$_1$** : *browse$_1$* the ~ (The user uses the I.)
(↔ fr: **Internet$_1$** : *naviguer$_1$* dans ~ (L'utilisateur utilise I))
*Both collocations are encoded with the* **Real$_1$** *LF.*

**Web$_1$**: *browse$_1$* the ~ (The internaut uses the W. )
(↔ fr: **Web$_1$** : *naviguer$_1$* dans le ~ (L'internaute utilise le W.))
*Both collocations are encoded with the* **Real$_1$** *LF.*

Figure 2: Search results for *browse*

This first strategy allows us to extract collocates that would not normally be considered as true equivalents, but in the contexts of the keyword there appear to be more valid solutions to suggest to users. For example, *move a mouse* is associated with three possible translations (*déplacer une souris, manipuler une souris, faire glisser une souris*), since all four collocations are encoded with the same LF, namely **Real$_1$**.

2. In the case of some paradigmatic relations, not only the presence of the same LF describing a lexical relation has to be verified, but also equivalence between the headwords pointed out by the LF in both languages. For example, the correct equivalents for co-hyponyms of *mouse* are shown in Figure 3. These relationships are encoded with Cf. (related meaning); however, since other pairs of terms are encoded with the same FL, equivalence between headword must be verified in order to extract the correct French equivalent.



**Related Meanings:**

**gamepad$_1$**: mouse$_1$ ( ≈ )
(↔ fr: **manette de jeu$_1$**: souris$_1$ ( ≈ ))
*Both relationships are encoded with Cf. (related meaning)*

**joystick$_1$**: mouse$_1$ ( ≈ )
(↔ fr: **joystick$_1$**: souris$_1$ ( ≈ ))
*Both relationships are encoded with Cf. (related meaning)*

**keyboard$_1$**: mouse$_1$ ( ≈ )
(↔ fr: **clavier$_1$**: souris$_1$ ( ≈ ))
*Both relationships are encoded with Cf. (related meaning)*

Figure 3: Search results for *mouse* and its co-hyponyms

### 4.3 Challenges

The strategies described in Section 4.2 work extremely well for the majority of lexical relationships that have been encoded up to now in the DiCoInfo. However, refinements will need to be made in order to extract valid equivalents in some cases. We identified two recurrent cases that will need to be dealt with separately.

a) Pointing to the right split actant: to describe some collocations, some actants – although having the same syntactic position – must be distinguished. For example, both *data* or *software* can be placed on a partition. However, collocates will differ if *data* is used (collocates are *copy*, *store* or *save*) or *software* is used (possible collocates are *copy* or *install*). For the time being, all these collocations are encoded with the LF **Labreal**$_{12}$. A further specification of the second actant of *partition* needs to be added in order to separate collocates that combine with *data* and those combine with *software*.

b) Distinguishing separate senses: Another problem is caused by the fact that some collocations are described with the same LF although collocates themselves differ in meaning. For example, the French collocations *cliquer sur … avec une souris* and *faire glisser … avec une souris* are both encoded with the LF **Labreal**$_{12}$ since the collocates both denote typical uses of the mouse. While this is not a real problem from the point of view of the description of the French term, the automatic establishment of equivalence relationships will produce erroneous results: *click on … with a mouse* will be associated with the two French collocations although only one applies.

## 5 Concluding remarks

We used the DiCoInfo as a case in point to bring a new perspective on the relationship between data encoding and structuring in TBD's on the one hand, and user-needs adapted data access on the other. These two dimensions are intimately intertwined. In the case of translation, working on the DiCoInfo has taught us that advanced semantic and syntactic encoding is not only a key for more and better data content, but also for better access from the user's perspective.

In this specific work we exploited a rich semantic encoding to extract equivalents of semantically related terms and, especially collocations. The encoding and access methods are completely transparent, i.e. users will benefit from the results without being burdened with an overloaded of information.

This work also has an unforeseen advantage. Terminologist entering the data can use these new search functions to locate inconsistencies in the descriptions (gaps, slight differences between languages, etc.).

### 6 References

FrameNet. 2009, http://framenet.icsi.berkeley.edu/. Accessed 26 November 2009.

L'Homme, M.-C. 2005. « Sur la notion de terme ». *Meta* 50(4), pp. 1112-1132.

L'Homme, M.C. 2008. « Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés », *Traduire* 217, pp. 78-103.

L'Homme, M.C. and P. Leroyer (2009). "Combining the semantics of collocations with situation-driven search paths in specialized dictionaries." *Terminology* 15(2), pp. 258-283.

Mel’čuk, I., Clas, A. and Polguère, A. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve (Belgique): Duculot / Aupelf - UREF.

Mel’čuk, I., Clas et al. 1984-1999. *Dictionnaire explicatif et combinatoire du français contemporain*. Montréal: Presses de l’Université de Montréal.

Ruppenhofer, J. et al. 2006. FrameNet II: Extended Theory and Practice. http://framenet.icsi.berkeley.edu/. Accessed 15 November 2009.

Appendix

**Internet**$_1$, n

***Status***: 2

**Actantial structure:**

the Internet: ~ used by Agent{user 1} to act on patient{information, site}

**Linguistic realizations of actants:** ...

**Synonym(s):** Internet network

**Contexts:**

*Certainly the Internet is the most conspicuous example of computer networking, linking millions of computers around the world, but smaller networks play a role in information access on a daily basis.* (Source: HOW ETHERNET WORKS)

*As we've seen in our discussion of the Internet and similar networks, connecting an organization to the Internet provides a two-way flow of traffic.* (Source: CURTIN)

*Each day, thousands more people gain access to the Internet (upwards of 6 million users at recent estimates).* (Source: WEB)

**Lexical relations:**

**Related meanings**

| | |
|---|---|
| ≈ Generic | Network |
| ≈ Related | meaning (Cf) intranet$_1$ |
| ≈ Related | meaning (Cf) extranet$_1$ |

**Types of**

| | |
|---|---|
| Which operates at high speed | - connection |
| Which operates at high speed | broadband ~ |
| Which operates at high speed | high-speed ~ |

**Combinations**

| | |
|---|---|
| In the I. | in the ~ |
| The user starts using the I. | access$_2$ the ~ |
| The user starts using the I. | connect to the ~ |
| → NOUN | access to the ~ |
| → NOUN | connection to the ~ |
| The user uses the I. | browse$_1$ the ~ |
| The user stops using the I. | disconnect from the ~ |
| → NOUN | disconnection from the ~ |

**Others**

| | |
|---|---|
| Division | Web$_1$ |
| Instrument to use the I. | browser$_1$ |

***français***: Internet$_1$