Compiling specialized dictionaries differently: A brief overview of terminological projects at the Observatoire de linguistique Sens-Texte (OLST)

Marie-Claude L'Homme

Marie-Claude L'Homme is full professor at the Department of Linguistics and Translation of the University of Montreal, where she teaches terminology and computer tools for translation. She is also the director of the Observatoire de linguistique Sens-Texte (OLST), a research group whose main focus is the lexicon (general and specialized) and the various theoretical and methodological issues related to its modelling. Her main research interests are lexical semantics applied to terminology and corpus-based terminology. Along with researchers in computer science, information science, terminology and linguistics, she develops lexical databases related to the fields of computing, the Internet and the environment.
mc.lhomme@umontreal.ca

1. Introduction

The Observatoire de linguistique Sens-Texte (OLST) is a research group located at the Department of Linguistics and Translation of the University of Montreal. It was created in 1997 with the aim of grouping experts with different backgrounds (linguistics, didactics, information science, terminology) interested in the various aspects of the lexicon. Its main objectives are the following:

- Address theoretical issues that underlie the various and complex properties of the lexicon.
- Develop and disseminate monolingual and multilingual resources (lexical and terminological databases as well as text corpora).
- Devise methods to apply this work to various applications (language teaching and information science, for instance).

The OLST also provides a multidisciplinary environment for training in lexicology, lexicography, and terminology. Several ongoing projects allow students to acquire the knowledge necessary to understand the theoretical issues related to the lexicon and the methodology necessary to compile lexical and terminological resources. In many cases, MA and PhD. students carry out research projects that are directly related to one of the ongoing projects at the OLST.

I focus here on two terminological projects that are in my opinion representative of the kind of environment and training that provides the research group. These projects and many other resources can be found on the OLST website, www.olst.umontreal.ca.

2. DiCoInfo and DiCoEnviro

DiCoInfo[1] and DiCoEnviro[2] are two lexical databases containing English, French and Spanish terms that are related to the fields of computing and the Internet (DiCoInfo) and to the field of the environment (DiCoEnviro), more specifically climate change and renewable energies. The descriptions provided in both resources are original in the sense that they aim to highlight various linguistic properties of verbs, nouns, adjectives and adverbs that convey a meaning that can be linked to either computing and the Internet (e.g. *administrator*, *download*, *graphical*, *dynamically*) or to the environment (e.g. *biomass*, *reduce*, *green*, *globally*). As can be seen in Figures 1 and 2 with the entries devoted to *warm*, the following information is provided:

- Clearly defined semantic distinctions: Figure 1 shows that the inchoative and causative meanings of the verb *warm* appear in two different entries. In fact, an additional entry describes the adjective *warm*: WARM$_2$, adj: ~ Patient{climate 1). In DiCoInfo, polysemous lexical items are also described in separate entries. For instance, *download* appears in three different entries:
  DOWNLOAD$_1$, vt: **Agent**{user 1} ~ **Patient**{application, file 1} from **Source**{computer 1, network} to **Destination**{computer 1} (*download hwclear.exe from Hauppauge's website*)
  DOWNLOAD$_{1.1}$, n: ~ of **Patient**{application, file 1} from **Source**{computer 1, network} to **Destination**{computer 1} by **Agent**{user 1} (*a download that never finishes or a similar problem*)
  DOWNLOAD$_{1.2}$, n: ~ of **Patient**{application, file 1} used by **Agent**{user 1}(*the download will be an executable file*)
- A number (status 2, 1 or 0) indicates how advanced the writing of the entry is. Entries with a status 0 (only available in the French version of DiCoInfo) are the most complete. Entries with a status 2 contain valid information, but are lacking some lexical relations and definitions.
- The actantial structure of terms: actants (i.e. arguments) are specified with two different labels (semantic roles such as **Agent**, **Cause**, **Patient**) and the typical term that can instantiate an actant (between curly brackets).
- Linguistic realizations of actants (Figure 2): a list of terms that can be found in running text and that can instantiate actants.

warm $_{1a}$, vi

Statut : 2

Domaine : climate change

Structure actancielle : warm: {climate $_1$} ~

Réalisations linguistiques des actants

Contextes

- *In winter, polar regions will also warm much more significantly than low latitude regions.* (Source : 5CANADAICC)
- *Because land masses warm faster than oceans, warming over the North American continent can be expected to be greater than the global average.* (Source : 1CANADACCD)
- *The main concern at present is to determine how much the Earth will warm in the near future.* (Source : ENCY)

Liens lexicaux

français : **réchauffer** $_{1a}$
espagnol : **calentar** $_{1a}$

Rédacteur(s) : MEL MCLH
Date de mise à jour : 29/10/2009

---

warm $_{1b}$, vt

Statut : 2

Domaine : climate change

Structure actancielle : warm: {energy} or {change} ~ {climate $_1$}

Réalisations linguistiques des actants

Contextes

Liens lexicaux

espagnol : **calentar** $_{1b}$
français : **réchauffer** $_{1b}$

Figure 1: Entries WARM$_{1a}$ and WARM$_{1b}$ extracted from DiCoEnviro.

warm $_{1b}$, vt

Domaine : climate change

Structure actancielle : warm: {energy} or {change} ~ {climate $_1$}

Réalisations linguistiques des actants

natural_force
aerosol , carbone dioxide , energy , gas , water
cause
change , effect , greenhouse effect , perturbation
patient
air , atmosphere , climate $_1$, earth , troposphere $_1$

Contextes

Liens lexicaux

Rôles actanciels

Explication - terme typique	**Lexie reliée**
Contraires	
Antonym	cool $_{1b}$
Autres parties du discours et dérivés	
Noun	warming
The climate w.	warm $_{1a}$

Figure 2: Linguistic realizations of actants and lexical relations in WARM$_{1b}$ (DiCoEnviro).

- Contexts: a selection of sentences extracted from the corpora used by lexicographers to write the entries.
- Links to French and Spanish equivalents when available online.
- A list of lexical relations (Figure 2): other terms that share with the headword a semantic relation such as hypernymy, near synonymy, antonymy, collocation. In some entries, the list of lexical relations is quite long. For instance, in entries for nouns, collocates are listed.

In some of the entries, definitions are provided. We display them in entries labelled with a status 0.

3. Resources designed according to lexico-semantic frameworks

Both resources are based on lexico-semantic frameworks that were originally designed to account for the general lexicon but that can be adapted to specialized lexical units. The first framework is that provided by Explanatory Combinatorial Lexicology (ECL, Mel'čuk et al. 1995). In DiCoInfo and DiCoEnviro, lexicographers refer to ECL when making semantic distinctions, defining the actantial structure, and listing lexical relationships.

The second theoretical framework applied in our resources is Frame Semantics (Fillmore 1982) and more specifically its application in FrameNet (Ruppenhofer et al. 2010). A module was recently added to both DiCoInfo and DiCoEnviro that shows how the term appearing as the headword interacts with its participants (actants and circumstants) in running text. Figure 3 shows part of the annotations and the summary table for the term ATTACH$_1$.

Of course, some adaptations were made to both frameworks when applied to the description of specialized terms.

4. A corpus-based methodology

It can easily be inferred from what has been said up to now that the methodology devised to compile DiCoInfo and DiCoEnviro is

Attach File - Click on this button if you want to send this email with a *file* **ATTACHED** to *it*. [INTERNET 0 JP MCLH 22/02/2009]

You can **ATTACH** more than one *file* to a single email *message* [INTROCOMP 0 JP 22/02/2009]

Email documents are created as text files so in order to send a binary file or document via email, *it* must first **BE** encoded into a text format and then **ATTACHED** to the email text *message*. [INTROCOMP 0 JP 22/02/2009]

Both methods allow *users* to sign or **ATTACH** a digital *identification* to the email *message* which verifies, to the recipient, that the message is from the original person or organization and that the information wasn't tampered with in transit. [INTROCOMP 0 JP 22/02/2009]

Netscape Communicator indicates that a *file* **IS ATTACHED** to a *message* by displaying a paperclip icon in the message window. [WIREDGUIDE 0 JP 22/02/2009]

ATTACH 1		
Actants		
Patient	Object (NP) (8) Subject (NP) (5)	file (9) document (2) it identification
Destination	Complement (PP -to) (14)	message (11) posting email it

Figure 3: Annotated contexts in ATTACH$_1$ (DiCoInfo).

the radiative forcings of all the trace gases (CO2, CH4,	**ozone** (O3), etc.) and aerosols. The total forcing may be treated a
e: water vapour, carbon dioxide, methane, nitrous oxide,	**ozone**, and halocarbons. Climate Trends An analysis of temperature r
ccurring greenhouse gases: water vapour, carbon dioxide,	**ozone**, methane and nitrous oxide. The amount of radiative forcing c
by these trace components: water vapour, carbon dioxide,	**ozone** and methane. It is this dependence on the minor components th
ffects from changes in concentrations of carbon dioxide,	**ozone**, methane, water vapour, line-shaped contrails, and aerosols,
use gases include water vapour, carbon dioxide, methane,	**ozone** and nitrous oxide. The magnitude of the natural greenhouse ef
ated if other forcing factors such as solar irradiation,	**ozone** depletion and non-sulphate aerosols are added. Once the credi
ated if other forcing factors such as solar irradiation,	**ozone** depletion and non-sulphate aerosols are added. Once the credi
one concentrations are also spatially variable. Further,	**ozone** is not a directly emitted species, but rather it is formed ir
eric components that are rapidly recycled: water vapour,	**ozone** and aerosols (liquid or solid suspensions in the atmosphere).
oping countries (section 3.8 in the TS). 2. Halocarbons,	**ozone** depletion and climate change 2.1 How do the CFCs and their re
quality, and future availability issues. 2. Halocarbons,	**ozone** depletion and climate change 2.1 What are
f the cooling effect and of reducing ODS concentrations,	**ozone** is likely to increase over much of the stratosphere, but coul
hings, emissions of CH4 and pollutants (as noted below).	**Ozone** concentrations respond relatively quickly to changes in the e
e third most important greenhouse gas after CO2 and CH4.	**Ozone** is formed by photochemical reactions and its future change wi
etres higher (at about 17 to 20 km) in the stratosphere.	**Ozone** in the upper troposphere and lower stratosphere is expected t
ents because they deplete the stratospheric ozone layer.	**Ozone** in the troposphere (the lower part of the atmosphere) is anot
NOx lead to decreases in the stratospheric ozone layer.	**Ozone** precursor (NOx) residence times in these regions increase wit
everywhere and is independent of where emissions occur.	**Ozone** also differs from the other greenhouse gases in that it is nc
vities assuming the mid-range IPCC scenario (IS92a). 4.2	**Ozone** The NOx emissions from subsonic aircraft i
with reduced or negligible global warming potentials. 4	**Ozone** within this report refers to stratospheric ozone unless other
all compared to those for greenhouse gases. Explanation:	**Ozone** depletion allows more ultraviolet radiation to reach the lowe

Figure 4: Concordances for *ozone*.

heavily corpus-based. For each resource, corpora had to be compiled in English, French and Spanish according to criteria adapted to terminology work (Bowker and Pearson 2002). In addition, corpora are enriched periodically to ensure that they are up-to-date.

Lexicographers use an in-house concordancer for obtaining sentences in which relevant terms appear (Figure 4). Between 15 and 20 contexts are selected and placed in entries. Contexts are then annotated (cf. Section 3). Lexicographers find most of the information necessary to fill the data categories of the resources in corpora. However, in some cases, they must also refer to existing specialized dictionaries or experts to validate a piece of information that was found in the corpus or access information that is not readily available in running text.

5. A computer-assisted process

Nearly all steps required to compile entries in DiCoInfo and DiCoEnviro are computer-assisted. We already referred to the use of a concordancer for finding relevant contexts. However, many other stages are partly automated.

The selection of terms to include in the word list is carried out using a term extractor called TermoStat and designed by Drouin (2003). The programme provides a list of candidate terms found in corpora based on a statistical calculation designed to define the specificity of lexical item in a specialized corpus when compared to a corpus of a different nature. The most recent version of TermoStat has many other features that allow lexicographers to have different views on the data (grouping of terms according to their components, identification of potential actants, etc.).

In the DiCoInfo and DiCoEnviro projects, lexicographers use the list of candidate terms generated by TermoStat to make a first selection of terms that will appear in the word lists of each resource. Figure 5 shows some of the results obtained when submitting an English corpus of climate change texts to the program. *Climate*, *change*, and *emission* were identified as the most specific terms in the corpus and now appear in DiCoEnviro.

Terms are then placed in an XML structure containing all the labels for data categories that lexicographers will fill with information based on what can be found in the corpora and their own knowledge of the fields.

6. A step-by-step training process

In order to compile entries in DiCoInfo and DiCoEnviro, lexicographers must become acquainted with some principles of ECL and Frame Semantics, be comfortable with the different computer programs used throughout the process (term extractor, concordancer, XML structure, etc.) and acquire some confidence as to their own judgements and intuitions about the meaning of terms. All this knowledge cannot be acquired at once, especially in our environment where most people working within our projects are students in translation who have no prior practical training in lexicography. It must also be pointed out that students work part time on the projects.

In order to ease the learning curve, a step-by-step process was defined. It usually flows as follows:

a) Read a couple of in-house documents to get a general overview of the methodology.
b) Start collecting contexts and place them in relevant entries.[3] In doing so, lexicographers must make semantic distinctions and often create new entries to reflect the polysemous nature of some lexical items. Lexicographers can carry out this work for two to three weeks before they start adding information to other data categories.
c) Collect true synonyms and graphical variants: these can be found in existing reference works.
d) Define the actantial structure: when they become comfortable with semantic distinctions, lexicographers are asked to

The Observatoire de linguistique Sens-Texte (OLST), a research group located at Université de Montréal, began to cooperate with K Dictionaries (KD) in 2010. Four OLST members have contributed articles to this issue of KDN:

- Marie-Claude L'Homme gives a general overview of the work carried out at the OLST and focuses on two specific terminological projects: DiCoInfo, *Dictionnaire fondamental de l'informatique et de l'Internet*, and DiCoEnviro, a database with terms related to the field of the environment.
- Marie-Claude Demers presents a methodology for identifying lexical items and meanings specific to the fields of computing and the

▶

Lemmatized candidate term	Frequency	Weight	Variants
Climate	5020	235,0826658	climate___climates
Change	4556	188,691871	Changes
Emission	3049	182,802937	Emissions
Global	1667	128,6522177	Global
Temperature	1733	126,2666191	temperature___ temperatures
Model	1669	117,5733414	Models
Scenario	1228	114,4488449	Scenarios
Carbon	1268	114,0777962	Carbon
Greenhouse	1299	113,2565932	greenhouse___greenhouses
Gas	1654	111,8322141	Gases
Concentration	1088	102,5969814	Concentrations
Ocean	1041	101,0607715	Oceans
Impact	1215	98,62949399	Impacts
Atmosphere	1017	94,90083166	Atmosphere
Warming	853	94,18214054	Warming

Figure 5: Candidate terms extracted from a corpus of climate change.

start defining the actantial structures of terms. They normally start with verbs; then they define the structures for adjectives and nouns.

e) Annotate contexts according to the methodology defined within the FrameNet project: this requires a specific training in order to follow a strict annotation process that encodes semantic and syntactic information about terms and their participants.

f) Establish equivalence relationships between English, French and Spanish entries.

g) Collect other semantically related terms: antonyms, near synonyms, collocates, etc.

h) Encode lexical relations using the system of lexical functions provided by ECL (Mel'čuk et al. 1995).

During each step, students are asked to note all the questions they may have and these are discussed with a more experienced lexicographer. The latter also revises the data categories periodically and decides when an entry can be placed online.

7. And much more

The previous sections offered a quick overview of two terminological projects carried out at the OLST and show how the various linguistic properties of specialized terms can be described in online resources. Recently, our projects have attracted the interest of researchers working in areas different from terminology and of lexicographers working on the general lexicon. Since both resources rely on lexico-semantic frameworks and on a methodology that is very close to the ones used in standard lexicography, they seem to lend themselves to extensions that we did not foresee when we first started compiling them. This section presents some of these extensions.

It soon became obvious that our resources, especially in the case of DiCoInfo, which has a larger coverage than DiCoEnviro, could be compared to general lexical resources in order to find lexical items or meanings that are specific to specialized domains and that might be lacking in more general repositories. We first carried out a comparison with a lexical resource that is located at the OLST, namely the DiCo (Jousse and Polguère 2005)[4] that is also compiled according to the theoretical and methodological principles of ECL. This comparison led to a series of criteria that can be taken into account when adding meanings related to specialized fields of knowledge to a general resource (L'Homme and Polguère 2008). Another comparison was carried out between the English FrameNet and the English version of DiCoInfo to find meanings that could be missing in the general language resource. We found that most meanings covered in DiCoInfo could be considered for inclusion in FrameNet (Pimentel et al. 2012).

The "lexicographic" potential of DiCoInfo also led to another interesting project. With K Dictionaries, we devised a method for interchanging data taken from the English version of DiCoInfo and from the *Random House Kernerman Webster's College Dictionary* (RHKWCD). The wordlists of each resource were compared semi-automatically and this comparison led to the identification of missing lexical items or meanings in each resource (Camirand, herein; Demers, herein; Demers et al. 2012). First, terms such as *avatar*, *artificial intelligence* and *google* (verb) were added to DiCoInfo. Other terms or specific meanings, such as *arrow key*, *attach* (as in *attach a file to an email*) and *data-driven,* are considered for inclusion in RHKWCD. When introduced in each resource, entries must by written according to their respective style guides.

We recently designed an interface to access the varied general and specialized resources available at the OLST in order to provide a first glance at the different meanings that may have a lexical item in general language as well as in specific subject fields. The interface, called Olster[5], searches the various resources and extracts the entries and an example for each entry. Users can then access the entry as it appears in the resources.

The last project that will be presented here concerns the various adaptations that were made to DiCoInfo and DiCoEnviro to make them more accessible to users who do not have a background in ECL or Frame Semantics. The resources were initially designed as research environments allowing researchers and students to carry out different kinds of analyses on terminological data. However, some colleagues pointed out that some aspects of the presentation of the data in the online versions of the resources and the access to their various data categories could be modified so as to make them more compatible with specific user needs (L'Homme et al. 2012). This work led to changes in the display of information on-screen and to the addition of new search functions. For instance, users can now access translations of collocations (e.g. *send sth as an attachment -> envoyer qqch. en pièce jointe*). A browsing module was also introduced in the French version of DiCoInfo that allows users to access a collocate that expresses a specific meaning. More changes will be introduced in the near future.

▶

Internet. This project was carried out within the cooperation framework for the interchange of data between a general language dictionary, the *Random House Kernerman Webster's Collegiate Dictionary* (RHKWCD), and DiCoInfo, and is part of her MA dissertation.

- Geneviève Camirand explains how new lexical items and meanings are described when added to DiCoInfo, within the same framework of interchange with RHKWCD. The format of DiCoInfo is not entirely compatible with what is expected to be found in RHKWCD, but some parts of the entries can later be adapted to the specific requirements of a general dictionary.
- Suzanne Desgroseilliers describes the work she carried out in order to adapt the equivalents provided in an English-French dictionary to the French used in Québec. This project was undertaken within an internship program offered by KD, with the objective of providing a hands-on experience in lexicography.

Another angle of the cooperation between OLST and KD, touching also on the import of entries from RHKWCD to DiCoInfo, is presented in the paper by Demers et al. at Euralex 2012.

Special thanks to Marie-Claude L'Homme for helping with the publication of these articles.

Notes

[1] DiCoInfo can be accessed here: http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi/.

[2] DiCoEnviro can be accessed here: http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi/.

[3] In both projects, a preliminary list of terms is provided to lexicographers. When they start working on the projects, they are not asked to select terms themselves. However, in the course of their work, they may find that some terms are missing and add them to the word list.

[4] http://olst.ling.umontreal.ca/dicouebe/.

[5] Olster was designed by Benoît Robichaud, research assistant at the OLST: http://olst.ling.umontreal.ca/olster/.

References

Bowker, L. and Pearson, J. 2002. *Working with Specialised Languages: A Practical Guide to Using Corpora*: London: Routledge.

Demers, M.-C., Kernerman, I. and L'Homme, M.-C. 2012. Lexicographic interchange between a specialized and a general language dictionary. In *Euralex 2012 Proceedings*.

Drouin, P. 2003. Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology*, 9.1: 99-117.

Fillmore, C.J. 1982. Frame Semantics. In The Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin, 111-137.

Jousse, A.L. and Polguère, A. 2005. Le DiCo et sa version DiCouèbe. Document descriptif et manuel d'utilisation. Montréal: Observatoire de linguistique Sens-Texte (OLST).

L'Homme, M-.C. and Polguère, A. 2008. Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. Maniez, F. and Dury, P. (dir.), *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*. Lyon: Presses de l'Université de Lyon, 191-206.

L'Homme, M-.C., Robichaud, B. and Leroyer, P. 2012. Encoding collocations in DiCoInfo: from formal to user-friendly representations. In Granger, S. and Paquot, M. (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 209-234.

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

Pimentel, J., L'Homme, M.-C. and Laneville, M.E. .2012. General and Specialized Lexical Resources: A Study on the Potential of Combining Efforts to Enrich Formal Lexicons. *International Journal of Lexicography*, 25.2: 152-190.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C. and Scheffczyk, J. 2010. FrameNet II: Extended Theory and Practice. 12 November 2011. http://framenet.icsi.berkeley.edu/.

ASIALEX 2013 in Bali

The 8th International Conference of the Asian Association for Lexicography (AsiaLex) will be held in Bali, Indonesia, on 20-22 August 2013. Preparations are underway, the conference website Asialex2013.org is up and running, and the Call for Papers will be published shortly.

The conference theme is Lexicography and Dictionaries in the Information Age. It is expected to draw the participation of not only lexicographers but also linguists, translators, teachers, and others interested in lexicography and dictionaries. The featured speakers include Dr. Diah A. Arimbi (Indonesia), Prof. Henning Bergenholtz (Denmark), Dr. Adam Kilgarriff (UK), Prof. Robert Lew (Poland), and Prof. Yukio Tono (Japan). In addition to the parallel paper sessions, there will be special sessions for software developers and for publishers interested in presenting the innovative features of their products.

The conference program is organized by the Secretary of AsiaLex, Dr. Deny Kwary, and his team from Airlangga University, in Surabaya, with the professional assistance of a local event organizer. Bali is well-known for its beautiful beaches, so the conference venue will be a beach-front hotel, and the conference will include a rich social program in addition to the academic presentations.

***

It is both a challenge and an opportunity to conduct a lexicography conference in Indonesia. On the one hand, lexicography is under-developed. For example, the biggest national language dictionary, *Kamus Besar Bahasa Indonesia* (4th edition, 2008), contains only 41,250 lemmata and 48,799 sublemmata, and consists of a single volume with the total of 1,701 pages. On the other hand, Indonesia is a vast country, with a large number of languages, and many foreign visitors: it has over 260 million inhabitants, 746 local languages, and approximately 600,000 tourists coming from abroad every month (more than 200,000 to Bali). Therefore, dictionaries should play an important role here.

Most Indonesians still use old versions of dictionaries, and are not aware of the latest developments and innovative features of modern dictionaries. Dictionaries are not compiled with the help of corpora and dictionary writing systems, but are usually written by using simple word processors. Therefore, ASIALEX 2013 will offer special sessions for software developers to promote their dictionary software and for publishers to promote their up-to-date dictionaries.

Deny A. Kwary