Browsing the Terminological Structure of a Specialized Domain: A Method
Based on Lexical Functions and their Classification

Marie-Claude L’Homme, Benoît Robichaud, Nathalie Prévil
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7 CANADA
{mc.lhomme, benoit.robichaud, nathalie.previl}@umontreal.ca

Abstract

This paper describes a method for browsing relations between terms and unveiling the terminological structure of a specialized domain. The method consists in expanding a graph that takes as input the relations encoded in a multilingual terminological resource called the DiCoEnviro that contains terms in the field of the environment. In the DiCoEnviro, terminological relations are encoded using lexical functions (Mel’čuk et al. 1995) and further classified in families defined on the basis of the properties of relations. We seek to provide users with an explicit and intuitive representation of a wide variety of relations. We also make the most of the richness of the encoding, while implementing some graphical choices to make their interpretation as clear as possible for end users. The method is implemented in a tool called NeoVisual that provides access to more than 11,000 relations in English and 15,000 relations in French. Portuguese is also included and coverage in all languages will increase as new entries are added to the DiCoEnviro.

Keywords: terminological relation; paradigmatic relation; syntagmatic relation; graphical representation; lexical function

1. Introduction

In terminology, it is assumed that concepts in specialized domains and terms used to express them are part of a structure. Concepts or terms are defined according to the place they have in this structure and the relations they hold with others. This principle is taken for granted (at least partially) when describing terminological data and some resources represent this structure explicitly (thesauri, terminological knowledge bases, ontologies).¹

In this paper we describe a method for browsing relations between terms and gradually unveiling the terminological structure of a specialized domain. The method is based on the contents of a terminological resource called DiCoEnviro that contains terms linked to the domain of the environment. This specific work builds on three sources (that are further presented in Section 2):

- Terminological relations manually encoded in the DiCoEnviro with lexical functions (Mel’čuk et al. 1995);
- A method devised by Robichaud (2012) developed on the basis of relations encoded in another term base very similar in structure to the DiCoEnviro, i.e. the DiCoInfo (on computing and the Internet);
- A classification of terminological relations and natural language explanations superimposed on LFs to facilitate their consultation in a user-friendly version of our resources (L’Homme et al. 2012).

The method consists in expanding a graph that takes the relations from the DiCoEnviro as input and allow users to explore different parts of the terminological structure in which they are interested. Our method is targeted at end users of terminological resources. So we seek to provide them with an explicit and intuitive representation of a wide variety of relations. We want to make the most of the richness of the encoding of terminological relations in the DiCoEnviro, while implementing some graphical choices to make their interpretation as clear as possible for users. The graphical display should complement textual (dictionary-like) information and not replace it altogether. It is assumed that from the point of view of users a graph is a more suitable tool to visualize and navigate through the terminological structure of a domain especially since some of its features can be exploited by designers to highlight different properties of relations and the nodes that these relations link. However, we also believe that other forms of terminological information (definitions, annotated examples) are more easily obtained in textual format.

The paper is organized as follows. Section 2 provides some details on how terminological relations are encoded in the DiCoEnviro and previous choices that were made to facilitate their interpretation. Section 3 presents other resources that resort to graphs to display relations between lexical units, terms or concepts. Section 4 describes our method and the tool in which it is implemented. Finally, Section 5 draws some conclusions and mentions a few areas that we wish to explore in the future.

2. Relations and their encoding in the DiCoEnviro

The DiCoEnviro is an online terminological resource under construction that contains terms in different languages, i.e. English, French, Spanish and Portuguese. The resource differs from most terminological repositories in the sense that it encodes and describes terms viewed as lexical units rather than labels for concepts. In addition to relations commonly taken into consideration, such as hypernymy,
meronymy and exact synonymy, the DiCoEnviro describes a large variety of other relations, including paradigmatic and syntagmatic ones. Table 1 gives a short list of relations that appear in the resource along with examples for each.

Relations are manually encoded by terminologists using lexical functions, LFs (Mel’čuk et al. 1995; Polguère 2014). LFs can represent both paradigmatic and syntagmatic relations and can take into account up to three different linguistic properties: the syntactic structure of a collocation, the general and abstract meaning of a relation and, finally, the relation between related term and the argument structure of the keyword. Technically, an LF applies to a key word and yields one or a short list of values. Examples of LFs are given in Table 1.

From the point of view of encoding, LFs have several advantages. First, as was mentioned above, they take into account different properties of relations (syntactic, semantic and argument structure) and thus allow us to classify related terms accordingly. Furthermore, they are language-independent. Hence relations in different languages that have the same meaning are encoded with the same LF.

However, from the point of view of their presentation in resources, they raise some challenges. In the online version of the DiCoEnviro, terminological relations appear in an ordered list that can be quite extensive in some entries. Going through this list can soon become cumbersome. In previous work, we investigated different ways to alleviate the exploration of relations.

A first strategy – that was implemented early on in the resource – consists of grouping sets of relations in families based on their properties (paradigmatic vs. syntagmatic, related meaning vs. opposites, etc.). This is reflected in the way that relations are presented in Table 1: “Related meanings” and “Opposites” correspond to these families. ²

Then, since LFs are not well known and can be rather difficult to decipher, a second strategy consists of adding a natural language explanation and superimpose it on each LF (L’Homme et al. 2012, based on Mel’čuk and Polguère 2007). For instance, the LF Real, is often explained as follows: The ... uses a ~ (here the ‘~’ stands for the key word and ‘…’ stands for the argument that involved in this collocation). While LFs are language-independent, natural language explanations must be adapted to each language taken into consideration in the resource.

The strategies that we mentioned in the previous paragraphs are presented textually to the users of our resources. An alternative method was developed by Robichaud (2012) to show part of the relations graphically and implemented in a tool called DiCoInfo Visuel. The tool displays relations included in the families “Same meaning” and “Related meaning” (distributed slightly differently in the Visuel), “Opposites”, “Word families” in a graph. It also shows the typical arguments of predicative terms along with a label that corresponds to their semantic roles. Figure 1 shows how relations for the two meanings of the verb erode are displayed (yellow arrows are used for arguments; purple ones for word families).

² Later on we refined this first subdivision. In the French version of the DiColinfo (L’Homme et al. 2012; L’Homme and Jia 2015), we defined sets of lexical relations and organized them into a hierarchy of classes. In this work, we referred to this system when selecting proper LFs and explanations for lexical relations. Of course, some adaptations were necessary for relations in the field of the environment.

| Relation                          | Example(s)       | LF               |
|-----------------------------------|------------------|------------------|
| **Same meaning**                  |                  |                  |
| Exact synonymy, variants, symbols | carbon dioxide → | Syn              |
|                                   | carbonic acid gas| Symb             |
|                                   | → CO₂            |                  |
| **Related meaning**               |                  |                  |
| Near synonymy                     | agriculture →    | QSyn             |
|                                   | farming         |                  |
| Generic                           | carbon dioxide → | Gener            |
|                                   | gas             |                  |
| **Opposites**                     |                  |                  |
| Antonymy                          | sustainable →   | Anti1, Anti2,    |
|                                   | unsustainable   | Rev1, Rev2       |
| Contrastiveness                   | fauna → flora    | Contr            |
| Conversiveness                    | propel → run    | Conv0            |
| **Word families**                 |                  |                  |
| Same meaning, different POS       | abundant →      | Ao, So, Vs,      |
|                                   | abundance →     | Adv0             |
|                                   | warm → warming  |                  |
| Adjective with added meaning      | erode → eroding | A1, A2, Able1,   |
|                                   | erode → erodible| Able2, etc.      |
| **Linguistic realizations of arguments** |               |                  |
| Role label (e.g. Agent, Patient)  | warm → (Patient)| Encoded in       |
|                                   | climate,        | lists with role  |
|                                   | atmosphere,     | labels           |
|                                   | temperature,    |                  |
|                                   | ocean           |                  |
| **Types of**                      |                  |                  |
| Intensification                   | toxicity → high | Magn             |
| According to a location           | habitat →      | Hypo – Lieu      |
|                                   | terrestrial ~   |                  |
| **Combinations**                  |                  |                  |
| Typical use                       | habitat →       | Real, Fact,      |
|                                   | inhabit in a ~  | Labreal1:2       |
| Existence                         | species → ~    | Func1            |
| Creation                          | territory →    | Caus,Func0       |
| establish a ~                     |                  |                  |
| **Others**                        |                  |                  |
| Meronymy                          | Earth → continent | [Part], [Tot].   |
|                                   |                  | Mult, Sing       |
| Quantity                          | greenhouse gas → | Quant            |
|                                   | concentration of |                  |

Table 1: Examples of terminological relations in the DiCoEnviro

![Figure 1: Erode and related terms in DiCoEnviro Visuel (Robichaud 2012)](Image 306x127 to 541x257)
The method that we now propose builds on the strategies mentioned in this section. However, in contrast with Robichaud (2012) that focuses on a chosen subset of relations, we take them all into account assuming that they are equally relevant to the terminological structure. Our method is designed to make the most of the rich encoding of relations in our resource as well as their classification in broader families while ensuring that the information presented can be readily interpreted. Hence we seek to establish a balance between rich encoding and simplicity of presentation and interpretation.

Furthermore, we want to allow users to discover the terminological structure gradually starting from a specific term and the relations it holds with others. Then users can use this first substructure to unveil other subparts by browsing through new sets of terms and relations.

Finally, since the DiCoEnviro is multilingual, we want to allow users to move easily from one language to another when browsing relations and explore relations as they appear in different languages.

3. Graphical representations of lexical and terminological relations

Representing terminological relations in the form of graphs has become a standard method for displaying the various interconnections between lexical units, terms or concepts and for labelling relations they hold explicitly. In some cases, graphs are superimposed on textual representations. In what follows, we make a selection of resources based on their relevance for our own work, but many more resources resort to graphs. Corresponding illustrations appear in Figure 2.

BabelNet (Navigli and Ponzetto 2012), a large multilingual lexical resource, proposes a graphical view of disambiguated lexical items in which lexical, terminological and encyclopedic information can be displayed. Relations are subdivided into broad categories that can be easily identified since different colors differentiate them, such as semantically related form (e.g., habitat $\rightarrow$ nature), gloss related form (ecosystem $\rightarrow$ environment), derivationally related form (inhabit $\rightarrow$ livable), see also (inhabit $\rightarrow$ live out).

Other resources, such as the ontology Envo (2017) and the terminological knowledge base (EcoLexicon, Faber et al. 2016) for the environment, focus on conceptual relations. Envo proposes a directed graph that is superimposed on a textual taxonomy. EcoLexicon implements a directed graph with a variety of conceptual relations (e.g., “canopy” type of “habitat”; “ecology” studies “habitat”; “colonization” located at “habitat”; etc.). In addition to the graph, more information on concepts (definitions, illustrations, etc.) is provided.

Another lexical resource, The French Lexical Network, FLN (Polguère 2014) is, to our knowledge, the resource that proposes the most fine-grained labeling of relations between disambiguated lexical units. It represents lexical systems, a non-ontological model of the lexicon (Polguère
The labeling is based on lexical functions (Mel’čuk et al. 1995): the graphical representation is superimposed on a textual encoding of relations and both are made available. The representation itself exploits the mathematical properties of graphs to visualize small world networks (Gaume 2008).3

The types and granularity of relations taken into account vary from one resource to another. As was mentioned above, we focus on lexical relations (paradigmatic and syntagmatic) between terms associated with a specialized field of knowledge. Thus our resource can complement others that represent mainly relations between lexical units (BabelNet or the FLN) or relations between specialized concepts (Envo and EcoLexicon). As far as specialized resources are concerned, our resource is probably the one presenting the largest variety of relations.

Regarding granularity, our method is closest to the FLN in the sense that each relation is labeled with a lexical function that captures three different properties (syntactic, semantic and argumental). We wish to make some of these properties visually explicit in the graph while not overloading it with textual information. However, we also group relations into larger families in order to facilitate their discovery and consultation by users. In other resources, it seems that the price to pay for obtaining clarity of presentation is to present very broad semantic categories or ignore some important distinctions between different kinds of units (terms, encyclopedic knowledge and lexical units). Conversely, resources that favor granularity of description result in a presentation that is less user-friendly. Our method combines both strategies and allows users to enter the network of relations linked to a specific term starting from general categories and breaking down these categories into smaller and more specific pieces instead of the other way around.

The resources mentioned in this section use various strategies to make important distinctions visible in graphs. Different colors or shapes are used for different sets of nodes or edges that stand for different relations. Specific relations (or sets or relations) can also be displayed and others hidden or placed in the background on demand. Our method also uses strategies that are available with the objective of making the rich encoding of our relations explicit while remaining clear for users. We also want to allow users to navigate through a network of relations in an intuitive way.

4. Organization of terminological relations in the resource

We developed our method based on all the terminological relations encoded in the DiCoEnviro (Table 2) for English, French and Portuguese.4 It should be kept in mind that the resource is still under construction. Hence, the number of entries and relations are most likely to increase in each language. Furthermore, the coverage differs quite drastically from one language to another due to the fact that work is some languages started later than others.

| Language | Entries | Relations |
|----------|---------|-----------|
| English  | 982     | 11,942    |
| French   | 1,309   | 16,723    |
| Portuguese | 37     | 563       |

Table 2: Data in the DiCoEnviro taken into account in the graph (as of February 2018)

For each relation, the following information is encoded in an XML editor: the related term, a lexical function, an explanation to be displayed in the online version. As was said above, terminological relations are also placed by terminologists in more general families according to their formal or semantic properties or their relationship with the head word (see Table 1).

All relations are represented using lexical functions, most of them as defined in Mel’čuk et al. (1995). However, some adaptations were made to describe domain-specific relations as well as to account for some methodological choices. Three of these adaptations are explained in this section.

First, although a term can only share one semantic relation with another (disambiguated unit), some related terms can appear in different families. This occurs with related terms such as unsustainable with respect to sustainable: unsustainable is both an antonym and a term that shares a morphological relation with sustainable. Since morphological relations are extremely productive in specialized domains, we want our graph to account for both the semantic relation and the formal one. Hence, when encoding these term pairs, terminologists account for them with the relevant LF that describes their semantic relation with another term (unsustainable is an antonym of sustainable and vice versa), but they also indicate that they are related formally. An XML attribute is added to the description of the relation.

Secondly, most families account either for paradigmatic relations (Related meanings, Arguments, Opposites, etc.) or syntagmatic relations (Combinations). Still, some families — such as “Types of” and “Others” — contain both single-word terms and collocations since the same relation can produce two different kinds of linguistic structures. For instance, the idea of size can be realized in the form of a new term or in the form of a collocation: habitat → microhabitat; territory → large ~. These relations are represented with similar LFs and in the same family. However, when encoding collocations, equivalent, when available, is displayed but navigation in relations is not permitted in Spanish.

Moreover, microhabitat shares a morphological relation with habitat. For this particular related term, we need to account for its meaning (“small”) with respect to habitat, the paradigmatic and the morphological relationships. This is taken into account in our graph features (Figure 3).
terminologists also specify the way the collocate combines with the key word (e.g. habitat $\rightarrow$ large ~, occupy a ~).

Thirdly, some relations are domain-specific and could not be encoded with standard LFs, especially in the “Types of” family. Previous work was carried out to propose an encoding adapted to terms in the field of computing (L’Homme and Jia 2015). We used some of these proposals but needed to refine the encoding for environmental specific relations. We also subdivide exact synonyms into more specific categories: synonym (carbon dioxide $\rightarrow$ carbonic acid gas), symbol (carbon dioxide $\rightarrow$ CO$_2$), abbreviation (greenhouse gas $\rightarrow$ GHG); variant (microorganism $\rightarrow$ micro-organism); feminine (Fr. expert $\rightarrow$ experente).

4.1 Tool used to present relations

For some years now (and especially since the emergence of social networks), there has been a large number of computer programs available to generate (or draw) visual graphs representing networks. Within the present project, we selected the package Network from vis.js (de Jong et al. 2015-2017), which is a library of programmable functions that can describe and generate natively dynamic graphs directly (as JavaScript code) in a Web navigation browser.

It is worth mentioning that with this particular package we do not have to manage the orientation of the layout, or the placement of nodes. Apart from the aesthetic aspect of the drawing itself, the feature we preferred – since it was in line with our objectives – was the ability to easily interact with the graphs created by means of popups, clicks and drag features. In particular, this package supports dynamic changes in the shapes and colors of nodes and edges under certain conditions as users explore the network.

4.2 Organization of information in the graph

In our method, users start discovering the terminological structure with a disambiguated term (a choice also made in BabelNet®). Although polysemy is reduced when focusing on lexical items from the point of a specific subject field, there is still a sizable amount of polysemy in the DiCoEnviro. Furthermore, the listing of relations and their fine-grained encoding can only be carried out if a single meaning is considered.

In order to meet the objectives we defined (clarity of presentation and ease of interpretation), we made the following preliminary choices (Figure 3):

- Users first obtain a graph that displays the searched term and all other terms to which it is connected.
  - A search field appears on the left hand side of the screen. It is equipped with an auto-completion feature that shows which terms appear in the resource. The list contains terms in three of the languages taken into account in the DiCoEnviro (English, French and Portuguese).
  - The central node of the graph is the disambiguated term (a reminder of its meaning is given by means of its argument structure displayed on demand).
  - Users obtain a global view of the relations held by the term they searched and other terms.
- Only those terms that are directly connected to the searched term are presented.

Figure 3: Habitat in the NeoVisual

---

6 The FLN also accounts for disambiguated lexical units; moreover, it presents the relations between the different meanings of a polysemous item.
Related terms in a language are clustered into up to eight different families (displayed only if relevant for the search term).

Labels are provided for families and a different color is assigned to each.

Morphologically related terms can appear in a family labeled “Word family” colored in purple. However, they can also be encoded elsewhere. Hence, the corresponding node is colored in purple throughout the graph even if the related term is encoded in another family (Figures 3 and 4).

Within families, related terms are further distinguished with nodes of varying shapes:

- Nodes that correspond to paradigmatically related terms are shaped in squares.
- Nodes that correspond to syntagmatically related terms are shaped in circles.
- Nodes that correspond to instantiations of arguments are shaped in triangles.
- Nodes that correspond to equivalents in other languages are shaped in diamonds.

Users can explore more deeply the relations within a family:

- Selecting a family label results in having the relations in this family highlighted.
- Users can also zoom into a specific relation by selecting an edge.
- The label of the relation (an LF) and an explanation are provided in the form of a pop-up (Figure 5).

The navigation through the terminological structure is permitted with the following features:

- Nearly all nodes in a graph are disambiguated terms and can be used to generate another graph that has the properties mentioned above.
- A special feature is provided for synonyms and other lexical forms (abbreviations, variants) that convey the same meaning as the term represented by the central node. Selecting a synonym will highlight all edges leading to families indicating that relations are also valid for synonyms (Figure 6).

Users can keep track of the navigation history (as in BabelNet) and return back to a graph already visited with the same clickable mechanism. If users navigate between languages, the history allows

---

7 Since the DiCoEnviro is under construction, some nodes have not been completely disambiguated yet, but they will be once these terms are described in a separate entry of the resource. In these cases, nodes are hollow to indicate that the associated term is not fully disambiguated. This feature applies to related terms within a language but also to equivalents in other languages.
them to go back to a previous one. The history is displayed as long as users do not select a new term in the search window.

- Users can navigate in a different language by selecting an equivalent. This will result in generating graphs in different languages (Figure 7).

- Access to the textual version (the original DiCoEnviro) is permitted when clicking on the central node of the graph. Similarly, users can refer to the graph from each entry in English, French and Portuguese.

5. Conclusion and future work

The method described above is implemented in a tool called NeoVisual that is available at: http://olst.ling.umontreal.ca/dicoenviro/ neovisual/. This graphical tool takes manually encoded relations from a terminological resource and presents them in a way that allows users to first obtain a general picture of the set of relations shared by a specific term with others. Then, they can access more information about a set of relations – a family – or a specific relation by selecting an item in the graph. The tool also allows them to browse the terminological structure of a specialized domain, i.e. the environment. Since the NeoVisual takes the relations as they are encoded in the original resource, terminologists do not need to change any aspect of their methodology when they add information to the entries. The number or relations taken into account by the NeoVisual increases as more data is added to the DiCoEnviro. Furthermore, terms in new languages could be added to the resource and taken into account graphically. Spanish should be added to the tool in the near future.

Our objectives were to make the most of a rich encoding based on lexical functions while focusing on the clarity their presentation. The graphical functionalities allow users:

1. To organize terminological relations according to their general meaning by first displaying them in families;
2. To further explain their meaning by adding labels on specific edges;
3. To further classify terminological relations according to formal properties (paradigmatic, morphological, syntagmatic). All these properties can be visualized all at once without having to regenerate graphs for a specific term.

The navigation functionalities are also designed to facilitate navigation through relations between terms in a specific language and to move from one language to another. We believe that a graphical display of terminological relations nicely complements a textual (dictionary-like) resource. It has the potential of making distinctions between relations readily visible (paradigmatic vs. syntagmatic; different families, etc.). These distinctions would be much more difficult to account for in textual format. However, it should not replace the textual resource altogether since other forms of information are better represented in textual format (definitions, contexts, etc.).

Our resource could be used for teaching purposes in terminology or specialized translation. Our next step is to collect user feedback and assess to what extent graphical displays can be used as a source of information for understanding the meaning of terms and visualize their position in a terminological structure. Users could be asked to use resources that present relations in the form of text and graph; or compare graphs with different features.

Figure 7: Conserve in English, conserver in French and conservar in Portuguese and their terminological relations
6. Acknowledgements
This work is supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada. We would like to thank reviewers and colleagues for providing us with useful feedback on different versions of the NeoVisual.

7. Bibliographical References
Faber, P., León-Araúz, P. and Reimerink, A. (2016). EcoLexicon: New features and Challenges. In GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference, ed. by Kernerman, I., I. Kosem Trojina, S. Krek, and L. Trap-Jensen, 73-80. Portorož.
Gaume B. (2008). Mapping the Forms of Meaning in Small Worlds. Journal of Intelligent Systems 23: 848–862.
de Jong, J., A. de Mulder and Stellingwerff, L. (2015-2017). Vis.js: A dynamic, browser based visualization library. Netherlands: Almende B.V.
L’Homme, M.C. and Jia, Z. (2015). Classement des combinaisons lexicales spécialisées à base nominale dans un dictionnaire d’informatique. Cahiers de lexicologie 106, pp. 229-251.
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, pp. 211-236.
Mel’čuk, I., Clas, A., and Polguère, A. (1995). Introduction à la lexicologie explicative et combinatoire. Duculot : Louvain-la-Neuve.
Mel’čuk, I. and Polguère, A. (2007). Lexique actif du français. L’apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français. Bruxelles: De Boeck.
Navigli, R. and Ponzetto, S. (2012). BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. In Artificial Intelligence 193, Elsevier, pp. 217-250.
Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. In International Journal of Lexicography 27(4).
Robichaud, B. (2012). Logic Based Methods for Terminological Assessment. In N. Calzolari et al. (eds.). Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey

8. Language Resource References
BabelNet (http://babelnet.org/)
DiCoEnviro (http://olst.ling.umontreal.ca/dicoenviro)
DiCoInfo (http://olst.ling.umontreal.ca/dicoinfo)
DiCoInfo Visuel (http://olst.ling.umontreal.ca/dicoinfo/visuel.php)
EcoLexicon. Terminological knowledge base (http://ecolexicon.ugr.es/en/index.htm)
Environmental Ontology, ENVO (http://bioportal.bioontology.org/ontologies/ENVO)
Viz.js. A dynamic, browser-based visualization library (http://visjs.org).