

Capturing Syntactico-semantic Regularities among Terms: An Application of the FrameNet Methodology to Terminology

Marie-Claude L'Homme, Janine Pimentel

Observatoire de linguistique Sens-Texte (OLST)

Université de Montréal

{mc.lhomme;janine.pimentel}@umontreal.ca

Abstract

Terminological databases do not always provide detailed information on the linguistic behaviour of terms, although this is important for potential users such as translators or students. In this paper we describe a project that aims to fill this gap by proposing a method for annotating terms in sentences based on that developed within the FrameNet project (Ruppenhofer et al. 2010) and by implementing it in an online resource called *DiCoInfo*. We focus on the methodology we devised, and show with some preliminary results how similar actantial (i.e. argumental) structures can provide evidence for defining lexical relations in specific languages and capturing cross-linguistic equivalents. The paper argues that the syntactico-semantic annotation of the contexts in which the terms occur allows lexicographers to validate their intuitions concerning the linguistic behaviour of terms as well as interlinguistic relations between them. The syntactico-semantic annotation of contexts could, therefore, be considered a good starting point in terminology work that aims to describe the linguistic functioning of terms and offer a sounder basis to define interlinguistic relationships between terms that belong to different languages.

Keywords: syntactico-semantic annotation, terms, actantial structures, DiCoInfo, FrameNet

1. Introduction

An increasing number of lexical databases provide detailed information on the syntactic and semantic properties of lexical units (e.g. DiCouëbe 2012, FrameNet 2012). Unfortunately, this does not apply to terminological databases as few include rich linguistic information. The project described in this paper aims to fill this gap by proposing a methodology for annotating terms in sentences based on that developed within the FrameNet project (Ruppenhofer et al. 2010). The method is implemented in an online resource that contains terms related to the fields of computing and the Internet called *DiCoInfo*.

In this paper we focus on the methodology we devised, and show with some preliminary results how similar actantial (i.e. argumental) structures can provide evidence for defining lexical relations in specific languages and capturing cross-linguistic equivalents.

The paper is structured as follows. Section 2 lists the objectives of the project and states why we decided to add a module containing annotated contexts in specialized lexical databases. Section 3 briefly describes our methodology and mentions a few differences between our project and FrameNet. Section 4 shows how regularities between similar terms can be captured by means of their actantial structures. Section 5 discusses a few cases where the actantial structure of predicative terms can support decisions when defining cross-linguistic equivalence.

2. Why annotate terms?

The specialized database used in this project is called *DiCoInfo*¹ and contains terms related to the subject field

¹ The resource can be accessed at the following URL: <http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi>.

of computing and the Internet that belong to the parts of speech of noun, verb, adjective and adverb (e.g. *download*, *to browse*, *virtual*, *remotely*). The database is compiled according to the theoretical and methodological principles of Explanatory Combinatorial Lexicology, ECL (Mel'čuk et al. 1984-1999). It currently consists of three language versions: French (approximately 1,000 entries, including 15,000 lexical relationships), English (approximately 800 entries, including more than 4,500 relationships) and Spanish (approximately 100 entries are currently online).² *DiCoInfo* entries contain the following data categories: part of speech, actantial structure, linguistic realizations of actants (i.e. arguments), lexical relations, contexts, and equivalents in other languages (Figure 1).

Entries are written in an XML editor and transformed into HTML when posted on the Web. Due to the rather complex structure of entries, a series of search and browsing functions were added to allow users to access specific pieces of information (L'Homme et al. 2012, forthcoming). For instance, users can display only those parts of the entries in which they are interested, they can locate translations of terms directly without having to read the entire entries; they can also access a term that is listed as a lexical relationship, etc.

Up until recently, the descriptions lacked specific details on the combinatorial properties of predicative terms and their actants.³ A module containing annotated contexts was added to the entry (cf. Section 3). By adding this module, we wanted to:

² We also developed a version of *DiCoInfo* in Korean that included annotations (Bae et al., 2008). However, these entries are not online.

³ This is due to a decision made by the *DiCoInfo* team and not to a gap in ECL. The model does specify the syntactic properties of lexical units and their actants (in a data category called *Régime*).

download₁, vt Status: 2

Actantial structure: download: Agent([user₁](#)) ~ Patient(application, [file₁](#)) from Source([computer₁](#), network) to Destination([computer₁](#))

[Linguistic realizations of actants](#)

[Contexts](#)

[Lexical relations](#)

[Actantial roles](#)

Explanation - Typical term	Related term
Related Meanings	
Conversive	upload₁
Other Parts of Speech and Derivatives	
Noun	download_{1,1}
An application or a file that can be d.	downloadable₁
Result	download_{1,2}

Spanish: [descargar₁](#)
French: [télécharger₁](#)

Figure 1: Entry for *download*

1. Provide users of specialized lexical databases with complete descriptions of the syntactico-semantic properties of terms (displaying combinatorial possibilities between predicative terms, especially verbs, and their actants);
2. Build a resource that can be integrated into NLP applications (these applications could use the syntactic description to find relevant information in running text);
3. Provide lexicographers writing the entries with more data for validating their intuitions;
4. Provide some evidence on meanings that differ from the ones they convey in other specialized subject fields or in general language (Pimentel et al. forthcoming).

We also believe that these descriptions can guide lexicographers when capturing semantic relations among terms within a language or discover potential equivalents between different languages as will be shown below.

3. Annotation methodology

Our annotation methodology is based on that devised in the FrameNet project (FrameNet 2012; Ruppenhofer et al. 2010), but is adapted and simplified in order to meet our specific needs (details of our annotation rules are provided in L’Homme and Pimentel 2010).

The most important adaptation lies in the use of a bottom-up approach instead of a chiefly top-down approach. According to Fillmore et al. (2003), FrameNet lexicographers first identify and define the frames as well as their Frame elements (FEs), then they

make a list of relevant LUs evoking the frames on which they are working, and, finally, they validate their intuitions based on corpora. In contrast, DiCoInfo lexicographers first select terms by means of a technique of corpus comparison (Drouin 2003) and by applying lexico-semantic criteria (L’Homme 2004), then they make semantic distinctions and, finally, they collect and annotate contexts.

As a result, FrameNet and DiCoInfo lexicographers follow different procedures when they select contexts for annotation. Typically, the former (semi-automatically) choose contexts from the British National Corpus (BNC) that illustrate a given syntactic pattern of the target LU and then proceed to annotate them. Therefore, the number of contexts annotated in FrameNet may vary depending on the number of syntactic patterns admitted by a given LU. In contrast, for each term, DiCoInfo lexicographers annotate between 15 and 20 contexts from a corpus containing various texts on the subject fields of computing and the Internet. When selecting contexts, lexicographers try to find those that reflect the different syntactic and combinatorial patterns of the term together with its actants: active vs. passive; different positions of actants; actants realized or not, etc. It is assumed that this number of contexts is sufficient to capture most combinatorial patterns of terms and their actants in specialized corpora.

Up to now, approximately 500 verbs have been annotated in French, 300 verbs in English, and we have started annotating Spanish verbs, whereas FrameNet lexicographers have annotated about 10,000 LUs belonging to all parts of speech.

There also some terminological and theoretical differences in the annotation methodologies adopted in FrameNet and in DiCoInfo. In FrameNet, participants are called *Frame elements* (FEs) and are subdivided into core and non-core FEs. Core FEs correspond to elements that are conceptually necessary for the frame, whereas non-core FEs are conceptually independent of the frames they accompany. DiCoInfo uses the terminological distinction introduced by Mel'čuk (2004) between actants (participants in the situation denoted by the lexical unit) and circumstants (that can appear in sentences but that are not part of the lexical unit's meaning).

In FrameNet, most core FEs are unique across frames, such as *Sound_source* and *Sound* in the [Make_noise] frame, and even if some FEs such as *Agent* are used across frames, they will probably be defined differently depending on the context. In DiCoInfo, the labels are reminiscent of the version presented in Fillmore (1968) and they apply to a large number of terms within our database (e.g. *Agent*, *Patient*, *Destination*).

Furthermore, there are some distinctions that FrameNet lexicographers make that are not taken into consideration in DiCoInfo. In FrameNet, non-core FEs are subdivided into peripheral or extra-thematic. Peripheral FEs situate events in space (e.g. *Place*) and time (e.g. *Time*), they describe how the event takes place (e.g. *Manner*), etc, and they only introduce events if these are part of the frame. In contrast, extra-thematic FEs introduce new events (other frames) against which the main event is situated. DiCoInfo lexicographers do not differentiate between circumstants.

In FrameNet, grammatical function and phrase type information are derived algorithmically by a chunk parser and instantaneously displayed on the screen, this requiring FrameNet lexicographers to review the results of the automatic grammatical function and phrase type tagging to make corrections, if necessary (Fillmore et al. 2003: 319). In DiCoInfo, the annotation is still performed manually but an automatic method was developed for annotation in French (Hadouche et al. 2011a). First results on French are quite encouraging and show that a lot of time can be gained when using the automated method (Hadouche et al. 2011b).

When compared to FrameNet, DiCoInfo uses a simplified system of syntactic annotation. In fact, at the syntactic level, other differences can be mentioned: FrameNet lexicographers introduce a label for a certain number of non-instantiated FEs; and in DiCoInfo, some indirect syntactic relationships are taken into account (for instance, an actant realized in the form of the subject of a modal verb and the syntactic subject of the annotated verb is annotated).

Finally, DiCoInfo does not provide users with tables

illustrating the valence patterns of terms, but the information on the semantic and syntactic properties of the terms and their participants show all the syntactic functions and groups in which participants can be found in sentences (Figure 2). To sum up, in DiCoInfo the annotated contexts indicate:

1. The predicative term in capital letters (DOWNLOAD);
 2. Participants and their nature (actants: *you, file, directory*; or circumstants: *easily*);
 3. Semantic roles of participants (*Agent, Patient, Destination, Source, Manner*);
 4. Syntactic function of the participant (subject, object, complement, modifier);
 5. Syntactic group of the participant (NP, PP, AdvP).
- By now **you** should be able to locate websites on the net, **DOWNLOAD files from these websites**.
 - **DOWNLOAD the file to your download directory**.
 - Alternatively, **the person in charge of the printer server** could **DOWNLOAD these files** and install them using the Additional drivers.
 - The best services let **you** upload and **DOWNLOAD files easily**.

DOWNLOAD		
Actants		
Agent	Subject (NP) (1) Indirect link (NP) (2)	You Person
Patient	Object (NP) (3)	file (3)
Destination	Complement (PP-to)	Directory
Source	Complement (PP-from)	Website
Circumstants		
Manner	Modifier (AdvP)	Easily

Figure 1: Annotated contexts for the term *download*

In DiCoInfo, the annotation is carried out in the XML version of the database. A specific schema was designed in order to assist lexicographers when selecting actantial roles, syntactic functions and syntactic groups (Figure 3 shows a sample of the annotation in XML). Although this schema is written in French because it was the first language described in DiCoInfo, it has been used in the other two languages. However, a number of adaptations to our annotation rules were necessary (for other adaptations made to English, refer to Pimentel and L'Homme 2011):

1. The linguistic metalanguage was translated (e.g. Eng. *Patient* -> Es. *Paciente*; Eng. *Destination* -> Es. *Destino*; Fr. *Modificateur* -> Eng. *Modifier*)

2. Auxiliary verbs were added in English and Spanish (e.g. Eng. *do*; Es. *estar*, *ser*)
3. We needed to take into account some modals

that are specific to some languages (e.g. Es. *tener que*)

```

<contexte source="ASSDELL" statut="0" annotateur="MCLH" mise-a-jour="2008-03-06">
  <contexte-texte> By now you should be able to locate websites on the net, DOWNLOAD
  files from these websites.
</contexte-texte>
By no
  <participant type="Act" role="Agent">
    <fonction-syntaxique nom="Indirect link">
      <groupe-syntaxique nom="NP">
        <realisation> you </realisation>
      </groupe-syntaxique>
    </fonction-syntaxique>
  </participant> should be able to locate websites on the net,
  <lexie-att> download </lexie-att>
  <participant type="Circ" role="Patient">
    <fonction-syntaxique nom="Object">
      <groupe-syntaxique nom="NP">
        <realisation lemme="file">files</realisation>
      </groupe-syntaxique>
    </fonction-syntaxique>
  </participant>
  <participant type="Act" role="Source">
    <fonction-syntaxique nom="Complement">
      <groupe-syntaxique nom="PP">
        from these <realisation lemme="website"> websites </realisation>
      </groupe-syntaxique>
    </fonction-syntaxique>
  </participant>

```

Figure 3: XML annotation of terms

4. Capturing regularities in the actantial structures of terms

There are many ways in which the annotation method described above can be of assistance to lexicographers. First, it helps them define much more systematically the actantial structures of predicative terms, such as verbs. In fact, generalizations about the number of actants, their nature (Agent, Patient, Instrument), and the typical terms that instantiate them are much easier to make based on a formal annotation of a representative number of contexts. In addition, as the identification of the actantial structures of terms are facilitated by this annotation method, lexicographers are also able to capture in a more systematic way differences and similarities between terms that are described in the terminological database. A few examples are given below:⁴

- Semantic distinctions. These are shown by the difference in the number of actants or by the different nature of actants. In the database, semantic distinctions are displayed in the form of separate entries:

connect(1): Agent ~ to Destination
(*A user connects to the Internet*)

connect(2): Agent ~ Patient to Destination
(*Connect the computer to a hub*)

For instance, the verb *to connect* appears in two different entries (showing that it can convey two separate meanings). The annotation of the contexts in which it occurs revealed that in some cases the verb had two actants and that these actants differ in nature, whereas in others it had three. The same applies to the verb *to program*.

program(3): Agent ~ Patient in Material
(*The application was programmed in Java*)

program(4): Agent ~ Patient
(*The Interface Card is programmed to read...*)

- Synonyms, near synonyms and certain antonyms. Similarities between actantial structures indicate that verbs can be synonyms, near synonyms, antonyms, etc. This kind of information can be viewed in the data category called Lexical relations.

⁴ Other similarities can also appear between terms that belong to different parts of speech. For example, a verb and its corresponding nominalization have the same number and type of actants, but these might appear in a different order.

connect(1): Agent ~ to Destination
(*A user connects to the Internet*)
log on(1): Agent ~ to Destination
(*A new person can log on to the PC*)

zip(1): Agent ~ Patient
(*You can zip a file*)
unzip(1): Agent ~ Patient
(*You can unzip a file*)

- Regular alternations. For instance, in causative-inchoative or agent-instrument alternations one actant is omitted and another one changes position. This information is specified by means of sub-senses.

boot(1a): Patient ~
(*Windows is booting*)
boot(1b): Agent ~ Patient
(*You should try booting the computer*)

print(1a): Instrument ~ Patient
(*The printer is printing the file*)
print(1b): Agent ~ Patient with Instrument
(*You can print the document on this laser printer*).

5. Capturing similarities and displaying differences between languages

Annotation is also extremely useful when establishing equivalence relations between predicative terms. In DiCoInfo, we assume that terms with identical actantial structures in different languages are quite probably equivalents, even though some differences may exist at the syntactic level (e.g. different choices of prepositions in each language). What is more, based on the actantial structures of terms as well as on the linguistic realizations of the actants, we observe that a term in language 1 can have more than one equivalent in language 2. These equivalents are near synonyms in language 2. For instance, the near synonyms *connect(1)* and *log on(1)* are both valid equivalents of the French term *connecter(1)* as well as of the Spanish term *conectar(1)*. Examples below show some near synonyms in French for the two meanings of the English verb *connect*:

Fr. connecter(1): Agent ~ à Destination
(*L'internaute se connecte à l'Internet*)
En. connect(1): Agent ~ to Destination
(*A user connects to the Internet*)
En. log on(1): Agent ~ to Destination
(*A new person can log on to the PC*)
Es. conectar(1): Agente ~ a Destino
(*El usuario se conecta a las gigantescas bases de datos*)
En. connect(2): Agent ~ Patient to Destination
(*Connect the computer to a hub*)
Fr. connecter(2): Agent ~ Patient à Destination
(*Connecter l'imprimante au serveur*)

Fr. brancher(1): Agent ~ Patient à Destination
(*Il faut brancher le modem au serveur*)

Sometimes, however, differences between the actantial structures of two candidate equivalents can indicate that there are some gaps in a language. This is, for instance, the case of *load(1a)* and *charger(1)*, which were not considered (perfect) equivalents: the English verb can be used intransitively and has two actants; the French verb can only be used transitively and has three actants.

Fr. load(1a): Patient ~ in Destination
(*The program is loading in memory*)
En. charger(1): Agent ~ Patient en Destination
(*L'utilisateur charge le programme en mémoire*)

Similarly, the English *upgrade* has three arguments and no true lexical equivalent was identified in French (terminological databases suggest *améliorer* or *mettre à niveau* as potential equivalents of *upgrade*, but French verbs only have two actants).

En. upgrade(1): Agent ~ Patient to Substitute
(*Manufacturers upgrade their Notebook to new technologies*)

Finally, similarly to Pimentel and L'Homme (2011), who concentrated on the French-English language pair, the comparison of the linguistic realizations of the actants of the French and English terms revealed that, although some terms are valid equivalents, the semantic nature of one of their actants may differ. This also applies to Spanish equivalents. For instance, in the contexts of the term *descargar(1)* the linguistic realizations associated with the actant Agent can either refer to animate (*usuario*) or inanimate entities (*servidor*, *navegador*), whereas in the contexts of the equivalent terms *download(1)* and *telecharger(1)*, the linguistic realizations associated with the actant Agent only refer to animate entities.

Our annotation method also reveals syntactic differences between otherwise valid equivalents that might not be as easily captured with more traditional methods (establish equivalence relationships based solely on semantic components). We mention two cases below.

Some verbs can have different syntactic behaviour. For example, *click* in English can be used both transitively and intransitively (*click the icon*, *click on the icon*).⁵ The two structures were found in the corpora. In French, however, *cliquer* can only be used intransitively (*cliquer sur l'icône*).

Some verbs see their actantial structures realized syntactically in a different order. For example, the English verb *search* and the French verb *chercher* have three

⁵ This case differs from *load_{1a}* and *charger₁* mentioned above. *Click* (transitive) and *click* (intransitive) are both valid equivalent for *cliquer* (intransitive).

actants. In the English actantial structure, the second and third actant can be inverted (*Someone searches information on the Web; someone searches the Web for information*). In French, the inversion is not allowed (*Quelqu'un cherche de l'information sur le Web; *Quelqu'un cherche (dans) le Web pour de l'information*).

6. Concluding remarks and future work

In this paper we showed that a sound methodology based on the annotation of terms in their contexts allows lexicographers to capture semantic regularities and better characterize similarities and differences between terms within a language as well as across languages. Although manual annotation is time-consuming, it is extremely useful and leads to more systematic descriptions.

However, the procedure described here would benefit from an automated method for checking the consistency of actantial structures within a specific language. The method could well rely on the annotations so as to gather and compare information on the actantial structures, these functioning as some sort of frames or interlingual representations (Boas 2005).

We envisage several adaptations from the method proposed in Boas (2005) to our project. As the semantic labels used to describe the actantial structures of terms are often to be found in many entries, the automated method should take into account the order of the actants as indicated in the actantial structures. For instance, near synonyms and specific types of antonyms should have the same actantial structure and actants should appear in the same order; nominalizations of verbs and verbs should have the same number of actants, but this order is allowed to differ.

With this method terms could be grouped together in a systematic way, which could help lexicographers better organize the meanings associated with a given subject field. By examining the actantial structures of terms in different languages, lexicographers can also validate equivalents and show users the differences between a term in language 1 and a term in language 2.

Other improvements to our method are also considered. First, a more user-friendly display of the annotated contexts could be implemented in the interface of DiCoInfo. In addition to listing all syntactic combinations of actants and predicative units found in contexts, it could provide a summary of possible syntactic realizations. Secondly, in a bilingual version, an automatic comparison of the syntactic behaviour of equivalents could be displayed, allowing users to visualize them in a single screen.

On a more theoretical level, a sounder definition of semantic roles for terminological purposes could be

provided based on the work carried out in our database. In a recent extension of this method to a terminological database containing terms related to the field of climate change called DiCoEnviro⁶, we noticed that some semantic roles were used much more frequently (e.g. Cause) and others needed to be introduced (e.g. Degree). This might lead to an assumption that some roles might be dependent on the nature of terms found in certain specialized domains.

7. Acknowledgments

The work described in this paper is supported by the Social Sciences and Humanities Research Council of Canada (SSHRC).

8. References

- Bae, H.S., L'Homme, M.-C. and Lapalme, G. (2008). Semantic Roles in Multilingual Terminological Descriptions: Application to French and Korean Contexts. *Multilingual and Comparative Perspectives in Specialized Language Resources. Proceedings of the Workshop. Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco
- Boas, H. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography*, 18(4), pp. 445--478.
- Le DiCou be. Dictionnaire en ligne de combinatoire du fran ais. (<http://olst.ling.umontreal.ca/dicouebe/>). Accessed 8 March 2012.
- Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1), pp. 99--115.
- Fillmore, C. J. (1968). The case for case. In E. Bach & R. T. Harns (Eds.), *Universals in linguistic Theory*. New York: Holt, Rinehard & Winston, pp. 1--88.
- Fillmore, C.J. (1982). Frame Semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin, pp. 111--137.
- Fillmore, C.; Petrucci, M.; Ruppenhofer, J. and Wright A. (2003). FrameNet in Action: The Case of Attaching. *International Journal of Lexicography*, 16(3), pp. 297--332.
- FrameNet (<https://framenet.icsi.berkeley.edu/fndrupal/>). Accessed 8 March 2012.
- Hadouche, F.; Lapalme, G. and L'Homme, M.C. (2011a). Attribution de r oles s emantiques aux actants des lexies verbales. In *Traitement automatique des langues TALN 2011*, 27 juin au 1er juillet 2011, Avignon.
- Hadouche, F.; Desgroseilliers, S.; Pimentel, J.; L'Homme M.-C. and Lapalme G. (2011b). Identification des participants de lexies pr edicatives :  valuation en performance et en temps d'un syst eme automatique. In *Actes de la 9e conf erence internationale Terminology and Artificial Intelligence (TIA'11)*, Paris, France.

⁶ The DiCoEnviro can be found at the following URL: http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi

- L'Homme, M.-C. (2004). *La terminologie : principes et techniques*, Montréal : Presses de l'Université de Montréal.
- L'Homme, M.-C. (2008). Le DiCoInfo. "Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire* 217, pp. 78--103.
- L'Homme M.C., Pimentel, J. (2010). *Guide d'annotation des contextes anglais du DiCoInfo*, Montréal : Observatoire de linguistique Sens-Texte.
- L'Homme. M.-C.; Robichaud B. and Leroyer P. (2012, forthcoming). Encoding collocations in DiCoInfo: From formal to user-friendly representations. In S. Granger & M. Paquot (Eds.). *Electronic Lexicography*. Oxford: Oxford University Press.
- Mel'čuk I., et al. (1984-1999). Dictionnaire explicatif et combinatoire du français contemporain. Recherche lexico-sémantiques I, II, III, IV, Montréal : Les Presses de l'Université de Montréal.
- Mel'čuk, I. (2004). Actants in semantics and syntax. I: actants in semantics. *Linguistics*, 42(1), pp 1--66.
- Pimentel, J., L'Homme M.C. (2011). Annotation syntaxico-sémantique de contextes spécialisés : application à la terminographie bilingue. In M. van Campenhoudt, T. Lino & R. Costa (Eds.), *Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face au défi de la diversité*, Paris : Édition des archives contemporaines/Agence universitaire de la francophonie , pp. 651--670.
- Pimentel, J.; L'Homme M.C. and Laneville M.E. (forthcoming). General and Specialized Lexical Resources: a Study on the Potential of Combining Efforts to Enrich Formal Lexicons, *International Journal of Lexicography*.
- Ruppenhofer, J.; Ellsworth M.; Petruck M.; Johnson C. and Scheffczyk J. (2010). FrameNet II: Extended Theory and Practice. <http://framenet.icsi.berkeley.edu/>. Accessed 8 March 2012.