Retrieving and Codifying Lexical Information in Process Oriented Terminology Management[1]

María Isabel Tercedor Sánchez
Clara Inés López Rodríguez
Facultad de Traducción e Interpretación
Departamento de Traducción e Interpretación. Buensuceso 11. Granada 18071, España.
Universidad de Granada.
itercedo@ugr.es; clarailr@ugr.es

Abstract
The emergence of new information media has had an impact on the working methods of lexicography and terminology as well as in the products obtained. Among the new media, knowledge bases are a valuable source that allows for information to be tailored to the needs of different users. We present several ways of codifying lexical and phraseological information in order to build a knowledge base on Coastal Engineering having a terminological tool in English, Spanish and German.

1 Introduction

The methodology of corpus linguistics and the multimodality of information have been of paramount significance to lexicographers and terminographers alike, influencing the way lexical patterns are identified and retrieved, in the case of corpus linguistics, and the presentation of information in Internet-based resources, in the case of multimodality. The availability of enormous amounts of lexical data prompts the issue of to what extent the lexicographer/user is able to grasp relevant information about a particular headword. Furthermore, the user wants multiple channels of information to be available at a click of the mouse.

Puertoterm is a research project on Coastal Engineering conceived to offer visual and textual information, a multilingual glossary in English, German and Spanish, and visual aids for the conceptual information contained. The objectives of the project are:

(a) to build a multilingual corpus of texts on the field of Coastal Engineering
(b) to build a dynamic, process oriented terminological database linked to a multimodal knowledge base
(c) to specify the relations and interactions of the Coastal Engineering Event (Faber et al 2005)

[1] This research is part of the project *PUERTOTERM: knowledge representation and the generation of terminological resources within the domain of Coastal Engineering*, BFF2003-04720, funded by the Spanish Ministry of Education.

(d) to follow a clear and concise definitional language for term entries following the approach of previous research projects (Faber et al 2001)

In this paper we describe the Puertoterm project, and the methodology followed for the retrieval of lexical and phraseological information, as well as the way it interacts with other parts of the knowledge base.

2 Building the corpus

Before compiling our corpus, we followed a top-down approach: a list of keywords was used as a starting point to identify more keywords, and the important information surrounding them. The list of keywords was obtained by comparing and compiling different sources of information – mainly glossaries – made available by the group of engineers participating in the project.

After a basic list was obtained, macrocategories were established and relations identified (see 3.1) in order to come to an organisational structure of the domain: the Coastal Engineering Event. The next step was to build a corpus of texts in the languages of the terminological resource, therefore adopting a bottom-up approach. This task gave us the chance to get acquainted with the basic concepts of the field. The corpus at its present state is described in the chart below:

	English	Spanish
Bytes	27,238,692	34,262,816
Tokens	4,435,525	5,075,774
Types	68,685	115,358

Figure 1. Composition of the Puertoterm corpus.

The texts were selected on the basis of their relation to the field of Coastal Engineering, and the following criteria:

(a) reliability of sources: texts dealing with coastal management issues were selected on the basis of their author/sender; therefore, texts published by official institutions at international, national and regional level were considered reliable. In the case of texts dealing with scientific issues – research and information – high impact magazines, prestigious encyclopaedic works and university textbooks were chosen.

(b) topicality: given the relevance of many of the macrocategories in the field (coastal management, sustainable development, hydrological constructions), the selection of texts necessarily had to follow the criterion of date.

(c) genre: texts were chosen following pragmatic criteria such as function and register, ranging from texts aimed at the general public to highly specialised texts such as technical reports aimed at the expert.

(d) geographical relevance: setting up a knowledge base implies aiming at a wide and heterogeneous audience. In the selection of texts we have covered a wide range of geographical origins, and have put special emphasis on geographical variants when codifying terminological information in the database.

Once the corpus had been compiled, we used concordances to extract relevant knowledge. More specifically, we identified paradigmatic relations in the form of hyponyms, meronyms, synonyms and antonyms, and codified them as related concepts in the database. We also looked for syntagmatic information on a keyword in order to offer the lexicographer/terminographer information about selection patterns.

3 Focussing on a specialised domain: a dynamic perspective

Terminological work has traditionally focussed on the organization of concepts and lexical units in a specialised domain. However, establishing conceptual and terminological limits in a subject field is a difficult task. Specialised domains interact among them often making up interdisciplines. Furthermore, establishing limits between specialised language and general language is far from easy since there are many units in general language that participate in specialised domains with a different nuance or sense, often given through the collocates appearing with a particular keyword. For these reasons, we have considered a Frame Semantics perspective (Fillmore 1985) as a valuable means for constructing a process oriented and dynamic representation of conceptual relations prior to codifying lexical information. Such a representation does necessarily have to relate categories within some type of general event structure. The notion of *frame* (Fillmore 1976) is applied in our project as a system of concepts interrelated in such a way that one concept evokes the entire system.

3.1 Implementing a Frame Semantics Approach in terminology management

In building a dynamic knowledge base, the *frame* notion can be a means for establishing links between concepts and clusters. In a terminological database, as in a frame network, classification is involved since these networks are divided into domains, the domains into frames, and the frames can go through several levels of specificity by using hierarchical inheritance. The data extracted from our corpus allowed us to work from a starting set of events and processes that we refer to as the Coastal Engineering Event (CEE) (Faber et al 2005).

The Coastal Engineering Event (CEE) is a dynamic process representation that is initiated by an agent (either natural or human), and which affects a specific kind of patient (a coastal entity), and produces a result. These macro-categories (AGENT ⇒ PROCESS ⇒ PATIENT/RESULT) are the concept roles characteristic of this specialized domain, and the CEE provides a model to represent their interrelationships. Additionally, there are peripheral categories which include INSTRUMENTS. If we consider the macro-category PROCESS, in the frame CONSTRUCTION, the agent will always be human. Other frames allow for human as well as natural agents (recharge of an aquifer). There are further nuances if we take a multilingual approach. One language may indicate a human or natural agent with only one lexical item, whereas a different language may have two different lexical items depending on the nature of the agent. This is the case of the Spanish headword **pantano,** corresponding not only to the English headwords **marsh** and **swamp** (indicating natural agent), but also to the word **reservoir** (artificial agent), as can be seen in the bilingual entry of the *Oxford Superlex.*

This sort of dynamic structure implies focussing on corpus data to further identify both multidimensionality (Bowker and Meyer 1993) and sense differentiation between two apparent synonyms through the scrutiny of the frame elements activated in the corpus.

4 Concordances in specialised languages

Concordance analysis is relevant to any terminological project as it gives clues about conceptual information as well as lexical co-occurrence patterns of a keyword. In our research project, extracting concordances has a fourfold objective:

(a) Extracting conceptual information (conceptual concordances): acquiring knowledge about the subject field, its relevant concepts and their relationships in the field.
(b) Knowing co-occurrence patterns in the specialised discourse (structural concordances)
(c) Knowing the selection patterns of verbs (verbal structural concordances)
(d) Understanding the different senses of a word: semantic prosody, metaphorical extensions and word sense disambiguation.

4.1 Concordances in the extraction of conceptual information and in knowledge representation

Collocational information on a keyword offers conceptual information about the place a concept occupies within the ontology. It specifically tells us about the characteristics of a concept as far as its place in a hierarchy or a meronymic structure and can help us to further identify frame elements that interact in the field.

Given the fact that the domain of Coastal Engineering is interdisciplinary, it is not surprising that the lexical items activated in it are multidimensional, showing different classification parameters. The use of corpora can shed light on the multidimensionality of concepts within a domain. If we focus on the IS-A relation, in other words, if we look for the different hyponyms derived from a basic concept (i.e. wind) in the corpus, we come to grips with the different perspectives under which a specific term can be seen, and we can infer the basic categories underlying the domain. As opposed to the information provided by dictionaries and encyclopaedias, concordances allow the identification of more parameters for classification (direction, height, speed, intensity, scale, etc.).

Filtering concordances makes it possible for the terminographer to fit a particular set to a classification parameter or frame element. In multidimensional representations, where a concept can be classified according to different criteria, this sort of organisation is of paramount importance. In Figure 2, we can see the concordances of the search item *viento** (wind) tagged for the classification parameters of: DIRECTION, HEIGHT, SPEED, INTENSITY, CONVERGENCE, POSSITIVE/NEGATIVE EFFECTS, SCALE, PLACE, HUMIDITY/TEMPERATURE, FREQUENCY, PREVALENCE.

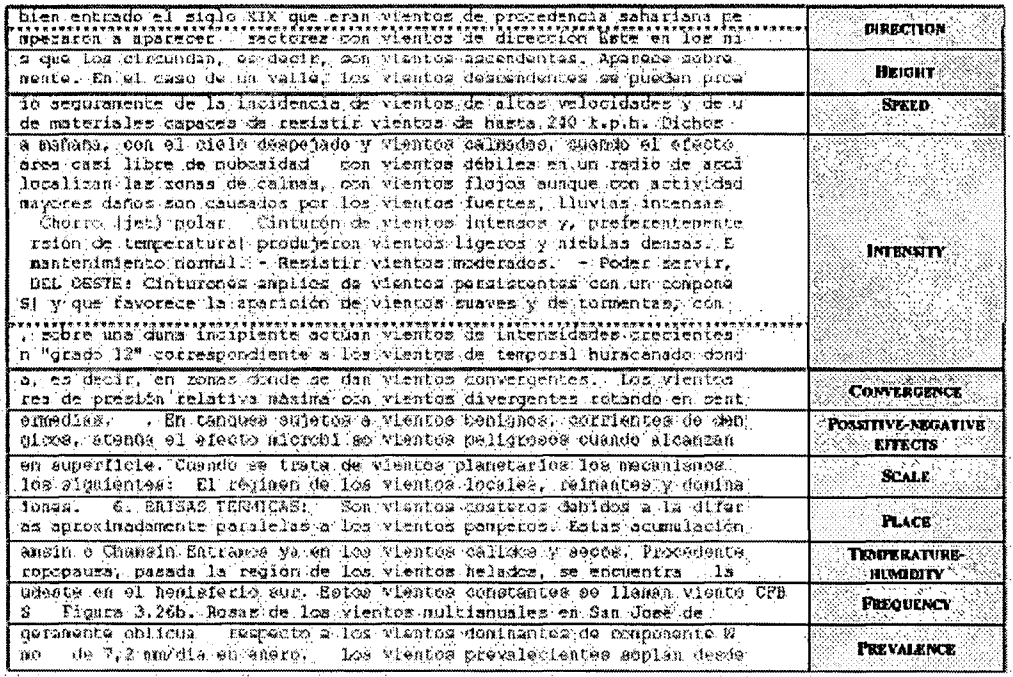

Figure 2. Filtered concordance displaying the multidimensionality of the term *viento* (wind).

Following a top-down approach in which multidimensionality parameters are identified and classification criteria established helps the terminographer to deal with the copious information concordances offer in an organised way.

4.2 Concordances to obtain collocational information

Concordances also show the different activations for each concept in real texts. If we take, for instance, the concept INTENSITY, most encyclopaedias will indicate that the strength of the wind is measured with a scale with 13 grades (Beaufort scale) and may include popular expressions to name these grades: *calma, ventolina, flojito, flojo, bonancible, fresquito, fresco,* [...] *temporal muy duro* and *temporal huracanado.*[2] The 324 hits for *viento* provided by the IATE database *(viento de mar, mar de viento, frescachón, viento fuerte,* etc.) offer no clue about classification parameters, essential to terminology management. However, if we take a look at the concordances pointing to the INTENSITY parameter (Figure 2), we can see more types of wind, and the elements phraseologically relevant for the keyword *viento,* for example, *calmados, débiles, intensos, fuertes, moderados, suaves* (adjectives) and *de intensidades crecientes, de temporal huracanado* (prepositional phrases).

[2] *Enciclonet* encyclopaedia.

4.3 Concordances to know the selection patterns of verbs

In our project, verbs are given a central role since they are key elements in a process oriented terminology management approach and allow a really dynamic representation of knowledge and lexical patterns (Faber et al 2005). Selection patterns in verbs tell us about the agent restriction of specific verbs; certain verbs restricting the agent to natural agents/ human agents. For instance, concordances show that the prototypical agent for the verb **to blow** in the domain of coastal engineering is WIND (Figure 3). Some morphological variants of this verb are also displayed. The concordances also point to relevant parameters defining the concept WIND (Figure 3, line 8): *strength, distance the wind blows (fecht) and the length of the gust (duration).* In Spanish, concordances show that these parameters are expressed with words such as *fuerza, velocidad, dirección* and *distancia* or *fetch.*

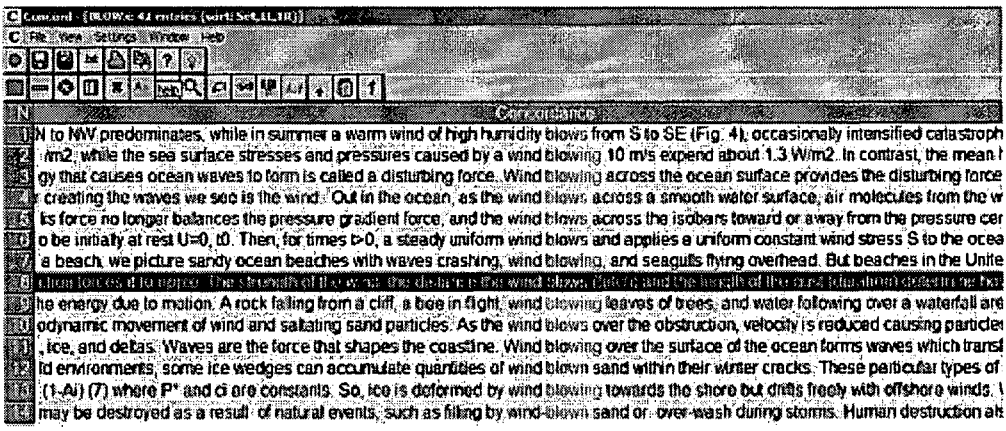

Figure 3. Concordance for *blow**

4.4 Concordances to understand the nuances in meaning of words

The study of the frames of verbs may help to disambiguate overlapping meanings and nuances that are carried in the manner an action takes place. This leads us to the study of the semantic prosody of words (Louw 1993). Noun phrases as direct objects of a verb and adverbs may tend to be negative (Atkins et al 2003: 272) or otherwise positive. In the previous concordance of the verb *to blow* (Fig. 4), the nouns, adjectives and adverbs accompanying it tend to take a negative nuance: *catastrophic, stresses and pressures, crashing, cracks, deformed...*

5 Conclusions

With our analysis we have focussed on the possibilities of concordance analysis for terminographical work. Concordances may be used to acquire expert knowledge and to understand the relevant concept in a subject field. Not only are syntagmatic structures retrieved through the analysis of collocates of a particular keyword, but also conceptual structures and

the interrelations between concepts. The identification of frame elements and their interrelations is necessary to codify lexical information and relations, and ultimately to build a knowledge base with a dynamic structure. The ongoing progress in the Puertoterm project will probably offer us much more insights into the structure and language of the domain of Coastal Engineering.

References

A. Dictionaries

Meiro, G. (2001-2005), *Enciclonet.* http://www.enciclonet.com [Access November 2005].
Rollin, N. (1999), *Oxford Superlex.* Oxford: Oxford University Press.
IATE. Terminological data bank of the European Institutions.
https://iate.cdt.eu.int/iatenew/consultation/search/sresults.jsp?PAGE=1. [Access November 2005].

B. Other Literature

Atkins, S., Fillmore, C. J., Hohnson, Chr. R. (2003), 'Lexicographic relevance: selecting information from corpus evidence'. *International Journal of Lexicography*, Vol 16, n 3. p

Bowker, L., Meyer, I. (1993) 'Beyond 'Textbook' Concept Systems: Handling Multidimensionality in a New Generation of Term Banks', in Schmitz, K.D. (ed.), TKE'93: *Terminology and Knowledge Engineering*, Frankfurt, Indeks Verlag, pp. 123-137.

Faber, P., López Rodríguez, C., and Tercedor Sánchez, M. I. (2001), 'Utilización de técnicas de corpus en la representación del conocimiento médico'. *Terminology* 7:2, pp. 167-197.

Faber, P., Márquez, C., Vega, M. (2005), 'Framing Terminology: A process-oriented approach', Paper presented at the symposium *For a Proactive Translatology commemorating the 50th anniversary of META, Translators' Journal.* University of Montreal, Quebec. Accepted for publication in META.

Fillmore, C. J. (1976), 'Frame semantics and the nature of language', in *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Volume 280, pp. 20-32.

Fillmore C. J. (1985), 'Frames and the semantics of understanding', *Quaderni di Semantica*, vol 6, pp. 222-253.

Louw, B. (1993), 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in Baker, M. et al. (eds.) *Text and Technology,* Amsterdam, Benjamins.