

John Benjamins Publishing Company



This is a contribution from *Terminology 14:2*
© 2008. John Benjamins Publishing Company

This electronic file may not be altered in any way.

The author(s) of this article is/are permitted to use this PDF file to generate printed copies to be used by way of offprints, for their personal use only.

Permission is granted by the publishers to post this file on a closed server which is accessible to members (students and staff) only of the author's/s' institute.

For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).

Please contact rights@benjamins.nl or consult our website: www.benjamins.com

Tables of Contents, abstracts and guidelines are available at www.benjamins.com

Integrating corpus data in dynamic knowledge bases

The Puertoterm project

Maribel Tercedor and Clara I. López-Rodríguez

Terminological information is a key element in the construction of a knowledge base. In order for a knowledge base to be useful to different users, terminological information should be extracted from corpora so as to reflect the different pragmatic nuances. PUERTOTERM is a knowledge base in the field of Coastal Engineering, which has made use of corpus information to develop terminological entries. It also includes contextual information in such a way that this information interacts with other elements of the knowledge base. We describe the methodology followed in the project regarding corpus design, retrieval of lexical information, conceptual organization of the domain of Coastal Engineering, and the elaboration of terminological entries.

Keywords: knowledge base, corpus linguistics, process-oriented terminology, Puertoterm project, concordances

1. Introduction

The methods of both terminology as well as lexicography can be enhanced by the use of knowledge bases, corpora and multimodal information sources. Knowledge bases (KB) are databases designed for knowledge management and allow for information to be tailored to different users, who can have access to different types of data according to their needs. The use of corpora facilitates the acquisition of knowledge about the subject field and the identification and retrieval of syntactic and lexical patterns. Furthermore, lexicographers and terminographers can benefit from multimodal information, that is to say, the presentation of information in internet-based resources, which provides the user with multiple channels of information such as text, image, sound and animation.

This study has been carried out within the PUERTOTERM project,¹ a R&D 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 product is a knowledge base (KB)² aimed at offering different users a vision of (1) the conceptual organization of the domain of Coastal Engineering; (2) terminological information proper, such as definitions, contexts, concordances; and (3) multimedia information such as videos, images. Such a knowledge base has two different interfaces: a researcher interface where knowledge is stored, and a final user interface based on MindMap technology, that will be available online. For the purposes of knowledge management, information is classified and stored into four sections which are coherent with the three aims described above: domains and relations, terms, and resources (see 2.5).

Since most of the information contained in the product is activated by the users according to their information needs, the target users range from technical writers to experts, translators, and the general public, who can benefit from the multidimensional and multimedia structure of the KB, the multilingual nature of the information and the different levels of detail in the information contained. At present, the database contains 2,838 concepts and 8,628 terms. The composition of the database as regards part of speech and language is as shown in Table 1.

Table 1. Part of speech of the terms in the PUERTOTERM database

	Spanish	English	German
NOUNS	1091	1160	1362
ADJECTIVES	103	98	46
VERBS	28	23	9
PROPER NOUNS	72	68	10
NOMINAL COMPOUNDS	2074	2069	415
TOTAL	3368	3418	1842

The main objective of the project is to build a dynamic, process-oriented terminological database (see Section 2.2) linked to a multimodal KB so that it allows intensional as well as extensional changes, in other words, it permits updates of information in real time from different remote systems and the interaction of users to activate relevant fields, specifying the underlying conceptual structure of the domain: its main concepts, the relations between concepts and the multidimensionality of the COASTAL ENGINEERING EVENT (Faber et al. 2005, Faber et al. 2006). In order to achieve this goal, it has been necessary:

- To build a multilingual corpus of texts on the field of Coastal Engineering.
- To extract knowledge from: (a) corpora (Faber et al. 2001; Pérez Hernández 2002; López Rodríguez 2003), (b) dictionaries and terminological databases, (c) exchanges with experts on coastal engineering, and (d) multimedia material (images, animations).

- 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 KB.

2. Methodology

In PUERTOTERM, we combine a bottom-up approach (extracting terms from corpora and placing them in a conceptual framework representing the field of coastal engineering) with a top-down approach (elaborating a list of domains and sub-domains where the different lexical units will fit). Our methodology covers the following cyclical, rather than sequential stages:

- Elaborating an initial list of keywords that is used as a starting point to identify more keywords, and the important information surrounding them.
- Compiling a multilingual corpus in English, Spanish and German.
- Organising the structure of the domain of coastal engineering with a conceptual frame representing the main processes involved: the COASTAL ENGINEERING EVENT (Faber et al. 2005). This conceptual representation explains all the concepts within this domain in terms of AGENTS, PROCESSES, PATIENTS, RESULTS, INSTRUMENTS and DESCRIPTIONS. Besides, this conceptual frame clearly indicates that specialized knowledge is embedded within our general knowledge about the world. To come up with these macrocategories we have applied both terminographic hierarchies as understood by the Functional Lexematic Model of Martín Mingorance³ (1984, 1989, 1995; Faber and Mairal Usón 1999; Faber et al. 2001, López Rodríguez 2007) and a Frame Semantics perspective (Fillmore et al. 2003).
- Using concordances in the extraction of conceptual information, in knowledge representation, and in the identification of syntactic and lexical patterns.
- Elaborating terminological entries with links to conceptual maps around the terminological entry.

2.1 Identifying the keywords of the domain

The list of keywords was obtained by comparing and compiling different sources of information — mainly glossaries and thesauri available in the internet, and sources made available by the group of engineers participating in the project. The quality of internet sources used in the project varied greatly as detailed in Faber

N	VWord	Freq.	%	Lemmas
1	AGUA	30.200	0,68	aguas(8402)
2	ZONA	8.693	0,20	zonas(4261)
3	SUELO	8.358	0,19	suelos(3262)
4	FORMA	7.154	0,16	formas(870),formación(1190)
5	SISTEMA	6.684	0,15	sistemas(2661)
6	RÍO	6.031	0,14	rios(2073),rios(37)
7	PARTE	5.919	0,13	partes(708)
8	MEDIO	5.846	0,13	medios(882)
9	GRAN	5.806	0,13	grandes(2371)
10	CASO	5.725	0,13	casos(1760)
11	NIVEL	5.372	0,12	niveles(1708)
12	USO	5.147	0,12	usos(1271)
13	MAYOR	5.002	0,11	
14	TIPO	4.979	0,11	tipos(1459)
15	ESTUDIO	4.860	0,11	estudios(1652)
16	CUENCA	4.842	0,11	cuencas(1452)
17	PROYECTO	4.600	0,10	proyectos(1547)
18	RECURSOS	4.585	0,10	recurso(1101)
19	SUPERFICIE	4.370	0,10	superficies(554)
20	AMBIENTAL	4.340	0,10	ambientales(1290)
21	FIGURA	4.254	0,10	figuras(239)
22	DESARROLLO	4.028	0,09	desarrollos(63)
23	CONDICIONES	3.969	0,09	condición(378)
24	GENERAL	3.758	0,08	generales(604)
25	TIEMPO	3.250	0,07	tiempos(261)
26	CALIDAD	3.212	0,07	calidades(19)
27	DATOS	3.209	0,07	dato(127)
28	ENERGÍA	3.085	0,07	energías(81)
29	TODO	2.964	0,07	
30	COSTA	2.867	0,06	costas(1142)

Figure 1. Lemmatised wordlist of the Spanish component of the PUERTOTERM corpus as extracted with the Wordlist application of WordSmith Tools

et al. (2006: 190). The material given to us by the experts had often a multimedia format and textual material had to be converted to *.txt to become part of the corpus.

The generation of lemmatised wordlists (Figure 1) using lexical analysis software (WordSmith Tools)⁴ also was useful in the identification of keywords which was used to get familiar with the basic concepts of the field. However, as shown in the most frequent keywords (*agua, suelo, río, cuenca*) initially the project concentrated on the field of River Hydrology, and texts were selected according to this field criteria. Since the focus was then extended to Coastal Engineering, the corpus is still being enlarged to cover this domain.

2.2 Compiling the corpus

In our multilingual corpora, the texts were selected on the basis of their relation to the field of Coastal Engineering and according to the following criteria:

- 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.
- Topicality: given the current 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 recent date, which in most cases corresponds to the 1980s onwards, except in the case of classical encyclopaedic works.
- 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. However, there is an imbalance with regard to the amount of texts dealing with a particular topic aimed at a particular audience: there are topics which have had a high impact in the online and printed press and therefore stand as more relevant in our corpus section aimed at the general public, whereas some expert topics remain in the specialised or semi-specialised level -such is the case of Earth Sciences.
- Geographical relevance: setting up a knowledge base implies aiming at a wide and heterogeneous audience, since information in the web is accessed by people from very different geographical areas. 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. Geographical variability has been achieved by compiling texts in Spanish, English and German published in different countries; this has allowed us to retrieve geographical variants for a particular concept (i.e. *bufadero* is the variant used in Spain for the concept BLOWHOLE, whereas in Mexico the corresponding term is *bufadera*).

The corpus at its present state is described in Table 2. Although the corpus also has a significant number of German texts, it is not comparable in size to the Spanish and English and for the sake of simplicity we are limiting our analysis here to English and Spanish. The English component of the corpus contains about 4.5 million words versus the Spanish one, over 5 million words. The difference in type/token ratio responds to the fact that Spanish has more inflection than English, especially

Table 2. Composition of the PUERTOTERM corpus

	English	Spanish
Bytes	27,238,692	34,262,816
Tokens	4,435,525	5,075,774
Types	68,685	115,558
Type/Token ratio	1.55	2.28
Standardised type/Token	36.81	39.92
Average word length	4.86	4.87
Sentences	283,104	179,926
Sentence length	15.46	27.23
Paragraphs	4,082	9,954
Paragraph length	380.59	275.75
1-letter words	362,132	457,108
2-letter words	735,010	1,252,230
3-letter words	770,488	680,333
4-letter words	547,517	337,202
5-letter words	430,692	403,467
6-letter words	320,458	338,073
7-letter words	332,193	374,221
8-letter words	304,791	333,340
9-letter words	227,400	272,957
10-letter words	170,332	235,171
11-letter words	112,352	163,024
12-letter words	58,323	94,491
13-letter words	37,179	64,476
14(+)-letter words	16,038	33,986

in verb conjugation and the inflection of adjectives (number and gender). Corpus data also highlight some of the assumptions made by linguists about rhetorical and terminological differences between languages. For instance, the assumption that Spanish texts contain longer sentences than English texts is validated by comparing average sentence length in Spanish and English (27.23 vs. 15.46). Moreover, Spanish texts are said to include more terms of Greco-Latin origin, as shown by the fact that the number of words with more than 10 letters is significantly higher in the Spanish component of the corpus than in the English one: for example, the number of 13-letter words in Spanish is 64,476 as opposed to 37,179 in English.

2.3 Organising the structure of the domain of Coastal Engineering: A Frame Semantics approach

Terminographical and lexicographical work starts with the delimitation of the object of description, be it the description of the relevant lexical units and their meaning in a specialised domain or in the general language for a particular purpose.

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, in PUERTOTERM we have considered a Frame Semantics perspective as a valuable means for constructing a process-oriented and dynamic representation of conceptual relations prior to codifying lexical information. The notion of *frame* is applied in our project as a system of concepts interrelated in such a way that one concept evokes the entire system:

The frame notion used in Frame Semantics can be traced to case frames (Fillmore 1976), which were said to characterize a small abstract situation in such a way that if one wished to understand the semantic structure of a verb it was necessary to understand the properties of the entire scene that it activated (Fillmore 1982: 115). A frame has been more broadly defined as any system of concepts related in such a way that one concept evokes the entire system. In building 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 evokes the entire system (Faber et al. 2005).⁵

The Frame Semantics perspective allows us to relate any category of the semantic network of Coastal Engineering to a general event structure which provides a framework for the basic processes and events that take place within this specialized field. In our project, the semantic network of Coastal Engineering is represented in a cognitive structure that we refer to as the Coastal Engineering Event (CEE) (Faber et al. 2005; Faber et al. 2006). The CEE (Figure 2) has been proposed after analysing the initial list of keywords, data extracted from our corpus, and after constant consultation with experts of the CEAMA (Andalusian Center for Environmental Studies).⁶

The CEE is a dynamic representation that is initiated by an agent⁷ (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/

The Coastal Engineering Event

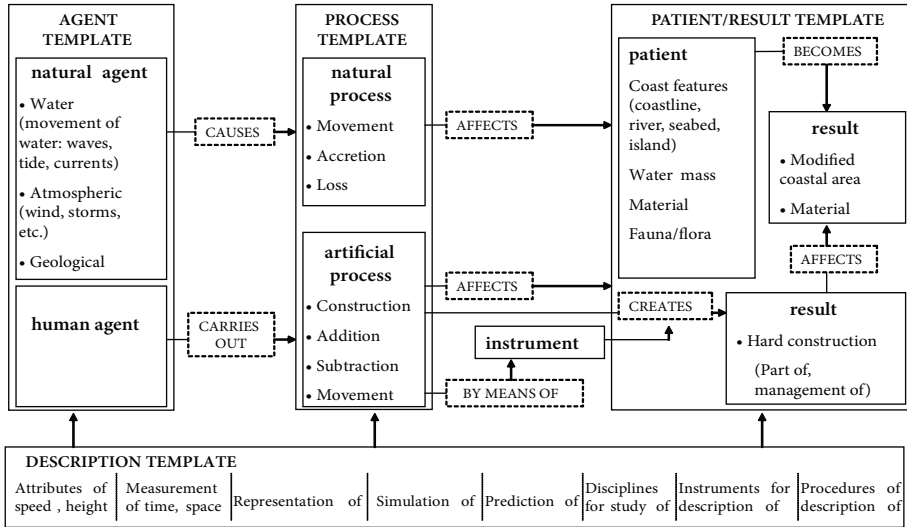


Figure 2. Conceptual representation of the COASTAL ENGINEERING EVENT (Faber et al. 2006)

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.

Each of the slots of the CEE is further developed into domains and subdomains that constitute the backbone of our knowledge base in the sense that any concept in the PUERTOTERM data base must fit into at least one of those domains. The complete list of domains and subdomains is accessible from the user interface of our KB (Figure 3).

Let us illustrate the way concepts are integrated in this knowledge structure with an example. If we consider concepts that activate the the subdomain ARTIFICIAL PROCESSES (B.2. in Figure 3) within the domain PROCESS (B), such as *coastal protection* or *dredge*, the agent will always be human. Other concepts such as RECHARGE OF AN AQUIFER allow for human as well as natural agents, therefore, this concept can be associated both with the subdomains of B.1.3. ADDITION (NATURAL PROCESS) and B.2.2. ADDITION (ARTIFICIAL PROCESS). 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).

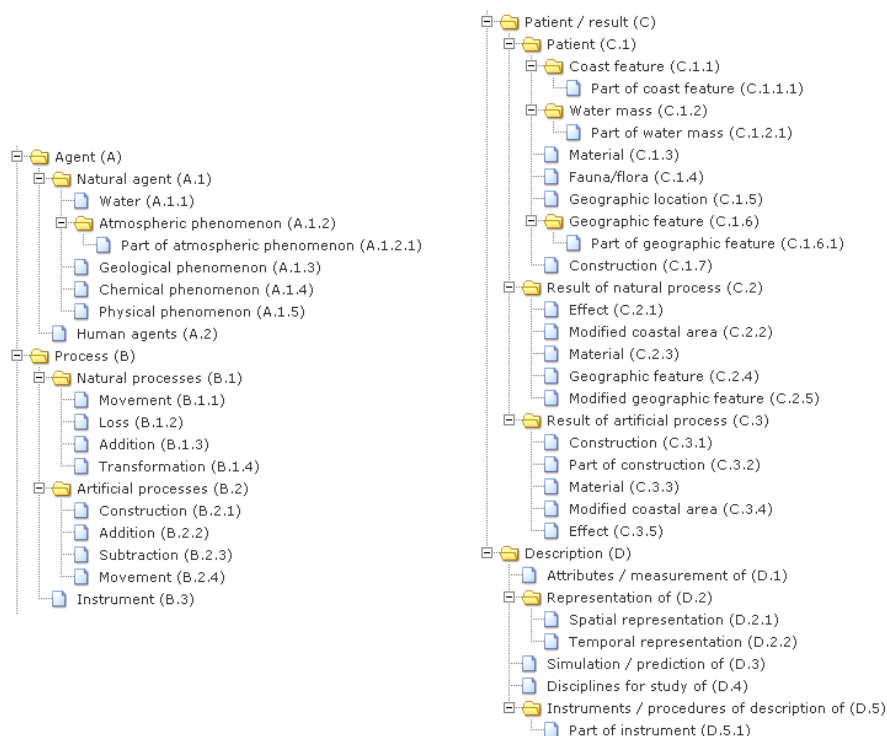


Figure 3. Representation of domains and subdomains in the PUERTOTERM knowledge base

The CEE is then a flexible frame that is fully compatible with several levels of specificity. More specific frames or sub-frames can fit into this frame, as it happens with the TIDE frame (Faber et al. 2006) or the WAVE frame (Prieto Velasco 2006; 2007: 362).

One of the advantages of this perspective is that concepts are organized around an action-environment interface (Barsalou 2003: 513; Faber et al. 2006). Such an interface is the result of concepts being considered as a representation to respond to the current needs of a situated action, rather than as an abstracted representation, and conceptual representations are seen as dynamic and contextualised (Barsalou 2003: 521). The name we give to this approach is Process-Oriented Terminology Management.

In our knowledge base, all concepts are linked with other concepts by means of relations.⁸ In fact, in the interface for researchers of our KB (created by Pedro Magaña Redondo), there is a tab devoted to RELATIONS. There we find, not only the habitual IS-A or PART-OF relation, but also AFFECTS, RESULT OF, MADE OF, HAS-FUNCTION, etc (Figure 4). These relations are displayed in the user interface of PUERTOTERM KB in a three-dimensional and dynamic representation (see Figure 6 in Section 2.5).

Concepto principal:	OLA
Relación:	resultado de
Concepto secundario:	CORRIENTE
<hr/>	
Concepto principal:	OLA
Relación:	resultado de
Concepto secundario:	RABIÓN
<hr/>	
Concepto principal:	OLA
Relación:	resultado de
Concepto secundario:	VIENTO
<hr/>	
Concepto principal:	OLA
Relación:	afecta a
Concepto secundario:	SISTEMA DE CORRIENTES LITORALES
<hr/>	
Concepto principal:	OLA DE TRASLACIÓN
Relación:	tipo de
Concepto secundario:	OLA
<hr/>	
Concepto principal:	OLA DE ELEMENTO

Figure 4. Some relations of the concept WAVE in our database (researchers' access)

This sort of dynamic structure implies focussing on corpus data to further identify relations between concepts, multidimensionality, and sense differentiation between two apparent synonyms through the scrutiny of the frame elements activated in the corpus.

2.4 Concordance analysis

Once the corpus had been compiled, and the general framework of the domain (with its relevant concepts and relations) had been identified, 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. The use of concordances will be further described in Section 3.

2.5 Elaborating terminological entries

In the researcher interface of our knowledge base we insert the entry in the general conceptual frame, the COASTAL ENGINEERING EVENT using the tabs RELATIONS and DOMAINS. For instance, the concept OLA (WAVE) can be located in a PROCESS template, in particular, in movement, but it can also be located in AGENT (physical phenomenon) and in a PATIENT/RESULT TEMPLATE (part of water mass), as Prieto Velasco (In press, 2007) points out. We also provide coherent and informative definitions enriched with visual resources. For example, in our definition of OLA (see Figure 5), the *definiens*, «oscilación» coincides with its superordinate concept. Besides, the definition activates other concepts and relations such as the ones present in Figure 4: VIENTO, CORRIENTE, RESULT OF, etc.

The display of information available to the user is shown in Figure 6 below. In the section *Términos* (terms) the terminological information is classified in each language; this includes definition, genre, part of speech, terminological variants, concordance lines and examples of usage. The section *Recursos* (resources) contains links to visual resources, which will allow both technical writers and translators to have a better understanding of the concept. The section *Dominios* (domains) places the concept in the relevant subdomains within the hierarchical structure of the Coastal Engineering Event, showing the multidimensional character of the concept WAVE: a wave can be profiled as a natural agent (physical phenomenon), as a process (movement), or as a result (part of water mass). The right frame of the screenshot shows the definitions and conceptual relations of «ola» as generated with MindMap technology.

Concepto	OLA
Definición	oscilación del nivel del mar en la interfase agua-aire generalmente producida por efecto del viento o de las corrientes.
Revisado	No
Autor	Miguel
Nota	

Términos Recursos Relaciones Dominios Mensajes

[+ Nuevo término](#)

Término:	ola	Más información
Idioma:	es	
Tipo de término:	término principal	

Término:	Welle	Más información
Idioma:	de	
Tipo de término:	término principal	

Término:	wave	Más información
Idioma:	en	
Tipo de término:	término principal	

Figure 5. The definition of WAVE in the researcher interface of the PUERTOTERM knowledge base



Figure 6. Display of information in the user interface of PUERTOTERM knowledge base

3. Concordances in Process-Oriented Terminology Management

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:

- Extracting conceptual information (conceptual concordances): acquiring knowledge about the subject field and obtaining definitions and encyclopaedic information.
- Knowing co-occurrence patterns in the specialised discourse (structural concordances).
- Knowing the selection patterns of verbs and pointing to verbs that collocate with certain keywords (verbal structural concordances).
- Understanding the different senses of a word: semantic prosody, metaphorical extensions and word sense disambiguation.

These concordances are not discrete types of concordances but point at discrete types of information that may be extracted after analysing and filtering concordance lines. These labels simply point to the type of information that is salient in the concordance.

3.1 Concordances in the extraction of conceptual, definitional and encyclopaedic information

Collocational information on a keyword offers conceptual information about the place a concept occupies within the ontology. It tells us about the characteristics of a concept as far as its place in a hierarchic structure, and can help us to further identify frame elements that interact in the field, allowing us to take multi-dimensionality into account. Concordances are therefore used to extract relevant knowledge through analysing the information surrounding a keyword, more specifically:

- Identifying paradigmatic relations in the form of synonyms, antonyms, and codifying them as related concepts in the database.
- Accessing syntagmatic information on a keyword, which offers the lexicographer/terminographer information about selection patterns.

In a knowledge base, concordance lines should also allow the user to access further encyclopaedic information. For instance, searching for Proper names around geographical features (i.e. Bay) gives access to local information that is not included in dictionaries but can be useful to suit special information needs. In these concordances lines (Figure 7) we learn about indigenous species (in bold italics and underlined) in a particular geographical area (in bold):

To elaborate good definitions the lexicographer can identify search structures, that is to say, linguistic patterns whose linguistic context is very informative for the identification of superordinate terms, synonyms, definitions, textual and

- 1 **ntroduction of exotic fish species**, as was found in **San Francisco Bay** ([Meng et al., 1994]), or by the dominance of fish species t
- 2 50 m) **benthic species**, known from the **Eastern Atlantic** (from the **Bay of Biscay** to Mauritania, the Azores, Madeira and the Canary
- 3 esh and Mridha [14] reported 103 **species of phytoplankton** in the **Bay of Bengal** including the North Indian Ocean. Belonging to the
- 4 ed **seven species of squids and two species of cuttlefish** from the **Bay of Bengal**. However, no report is available on their production and sto
- 5 ssociated **liver disease in two species of fish** from **San Francisco Bay and Bodega Bay**. *Ecotoxicology* 5 (1997), pp. 1-31. Vethaa
- 6 . K. Sasaki, Two new **species of Atrobucca (Sciaenidae)** from the **Bay of Bengal**. *Japanese Journal of Ichthyology* 42 3-4 (1995), pp
- 7 ment the use of vegetated and unvegetated habitats in **Rehoboth Bay** by **fish and blue crabs**. Many species of resident fish were f
- 8 eries database [101] reported as many as **629 fish species** in the **Bay of Bengal**. A complete list and description of the coastal and

Figure 7. Encyclopaedic information displayed in concordances around Bay

1 n (Figure 7f-2). **Absorption** is defined as a process in which solar radiation i
 2 discharge. A **recharge area** is defined as a portion of a drainage basin where
 3 ter table. A **discharge area** is defined as a portion of the drainage basin wher
 4 **channel cover factor**, CCH, is defined as the ratio of degradation from a chan
 5 are needed. A **unit channel** is defined as a channel of length $L = 1$ km and wid
 6 The **reflection coefficient** is defined as the reflected wave height divided by
 7 he **transmission coefficient** is defined as the ratio of the transmitted wave he
 9 o « SEDmx (days). A **dry day** is defined as a day with surface runoff less than
 10 n dry days, where a **dry day** is defined as a day with less than 0.1 mm of surfa
 11 the previous day. A **wet day** is defined as a day with 0.1 mm of rain or more. T

Figure 8. Concordance around the search structure “is defined as”

orthographical conventions, and translation strategies (López Rodríguez 2002).⁹ Search structures such as “is a”, “called” “kind* of”, “is defined as” help the terminographer to extract both the main concepts in the subject field and definitions to better understand those superordinate terms (Figure 8).

3.2 Concordances to obtain collocational information

Codifying a terminological entry for a particular keyword should be done in such a way that relevant co-occurrence patterns are made available to the potential user. Phraseological information can be analysed studying the collocational data surrounding a keyword. In Figure 9, we chose (in bold) the particular core elements which are phraseologically relevant for the keyword *ola* (WAVE):

Collocational information not only indicates syntactic co-occurring patterns but also sheds light on the conceptual characteristics of a keyword. In the case above, we can infer classification criteria for the concept OLA (WAVE) by determining the conceptual categories where collocates fit (PARTS OF THE WAVE, HEIGHT, STRENGTH).

- Concordance

steira), se ha medido ya un temporal con **altura** de ola significativa de
 os concluyeron con una determinación de altura de ola significativa de
 . Cada **período** se asoció a una serie de alturas de ola HS. En la Figura
 El **período T** es una característica constante de la Ola durante su
 s un arma de dos filos (18.6). V La altura de la ola depende de la
 ntras se llena, se vacía de golpe, produciendo una ola que limpia el 30
 ste entre una **cresta y un valle**, la **longitud** de la ola se refiere a la
 ón se ve muy influido por el valor de la altura de ola, tomando valores va clara-
 mente cómo los valores de la **potencia** de ola en el Golfo de, solo tenían una variable
 climática, la altura de ola, cuando todos e
 la importancia de otras variables que la altura de ola, tales como el
 nas de las variables medioambientales H (altura de ola), Ir (número de s que el
 coeficiente de variación de la altura de ola significativa

Figure 9. Concordance for *ola* (wave)

3.3 Verbal structural concordances

In PUERTOTERM, verbs are given a central role since they are key in a process oriented terminology management approach (Faber et al. 2005). The generation of verbal structural concordances sheds light not only to the verbs that usually collocate with relevant terms, but also to selection patterns of verbs to become terminological expressions. Figure 10 shows that the verbs that usually collocate with *marea* are: *arrastrar*, *bajar*, *descender*, *inundar*, etc.

Selection patterns in verbs tell us about the agent restriction of specific verbs, certain verbs restricting its 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 11). Some inflected forms¹⁰ of this verb are also displayed in Figure 11. The concordances also point to relevant parameters defining the concept WIND (Figure 11, line 8): *strength*, *distance the wind blows (fetch)* and *the length of the gust (duration)*. In Spanish, these parameters are expressed with words such as *fuerza*, *velocidad*, *dirección* and *distancia* or *fetch* (Figure 12).

Semantic prosodies (Louw 1993) are another object of study. Noun phrases as direct objects of a verb and adverbs complementing verbs may tend to be negative (Atkins et al. 2003: 272) or otherwise positive. In the concordances of the verbs *to blow/soplar* (Figures 11 and 12), the nouns, adjectives and adverbs accompanying it tend to take a negative nuance: *catastrophic*, *stresses and pressures*, *crashing*, *cracks*, *deformed*, *violentamente*, *más favorablemente*.

Arrastrar
bajar
descender
inundar
penetrar
subir

LÍNEAS DE CONCORDANCIA

1 de las olas, pero cuando sube la **marea** las olas pueden penetrar en el río. Por
 2 los detritos arrastrados por la **marea**, lo que conlleva una intensa actividad b
 3 dulce y salobre inundadas por la **marea**. I. Humedales intermareales arbolados; 4 a
 pared y esperar a que suba la **marea**. Marcar el nivel máximo del agua y medir 5 ayores
 en el fondo que cuando la **marea** disminuye, en tanto que en la superficie 6 de un mo-
 vimiento de subida de la **marea** (aguas frías profundas) a causa de un
 7 re los niveles alcanzados por la **marea**. Existe un límite (3 m) por debajo del
 8 ante el ascenso y descenso de la **marea**. Estuario del Regeg en Rabat(Marruecos)
 9 n de la máxima penetración de la **marea**, lo que facilita la tarea de la
 10 ea; al inicio de la bajada de la **marea** son aguas mixtas, saladas y a veces dulce
 11 provoca la subida y bajada de la **marea**. Estoscambios de altura del agua del mar

Figure 10. Verbal structural concordances around the word *marea* (tide)

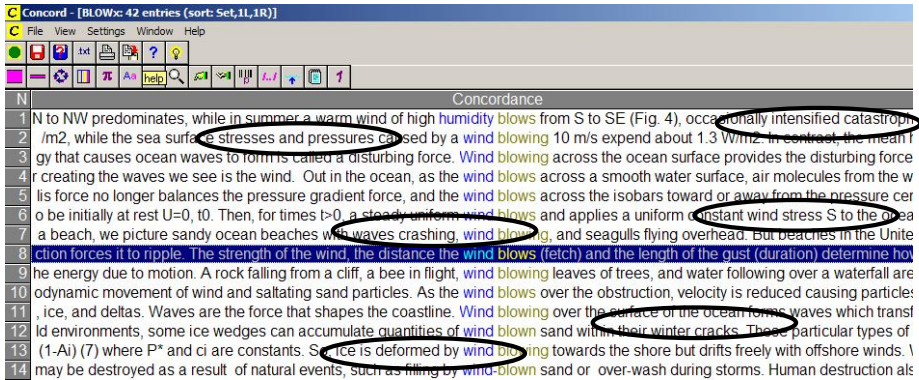


Figure 11. Concordance for morphological variants of *blow* in the PUERTOTERM corpus

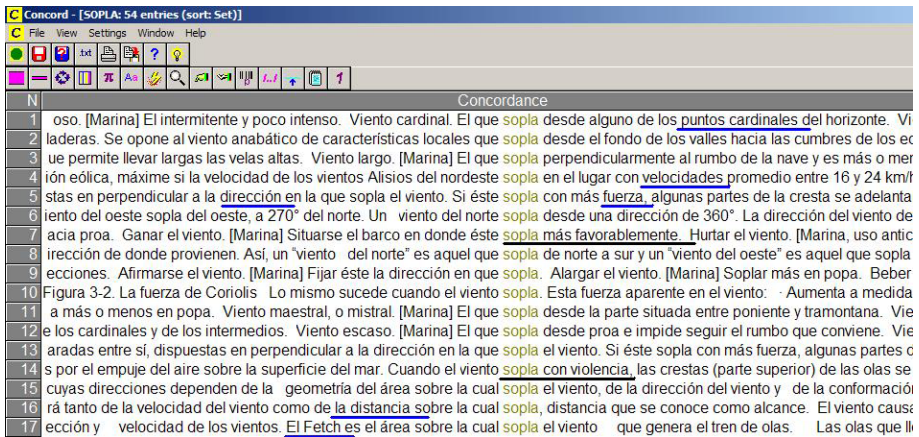


Figure 12. Concordance for morphological variants of *sopla** in the PUERTOTERM corpus

3.4 Concordances to disambiguate the senses of a word

Concordances can be used to determine polysemous senses of a particular term within a field. In order to do so, we have to analyse the place the concept takes in the system and the main characteristics making up core information. In the example below (Figure 13), the term *abastecimiento* is recurrent in the area of coastal engineering but has several different senses that are made visible through the syntagmatic structures co-occurring with it. Only the cases in bold are lexical structures relevant to the frame RECHARGE (OF AN AQUIFER), whereas the rest of instances are related to the frame of WATER/ENERGY SUPPLY, and appears mainly in texts produced by public bodies.

The same polysemous nature can be seen codified in the terminological data contained in the IATE database of the European Union. However, the headword *abastecimiento* will not be retrieved in a search from English into Spanish of the

Concordance

1- UBICACIÓN – AREA DE ESTUDIO El abastecimiento de agua para la población se realiza el 100% del agua utilizada para abastecimiento. Entre estos últimos destacan los iem N Estudio Hidrogeológico para el abastecimiento a Motril (Granada). Diputación rovi Ctura energética. Elementos para abastecimiento de energía a la población y a las tificados – Consolidación del abastecimiento de agua potable para los grandes úcle Porte 5.4. **Hidroenergía** 5.5. **Abastecimiento de agua** para consumo humano 5.6. PROPUESTOS 7. 1 Un **depósito de abastecimiento** tiene la superficie libre a la cota . VIGENCIA 0. Introducción El abastecimiento de agua para uso y consumo humano os a embalses utilizados para el abastecimiento de aguas potables, como por ejemplo bre todo, a través de **sondeos de abastecimiento** y con fines agrícolas. En el esta , a la producción de energía, abastecimiento municipal e industrial, control a EMASESA (Empresa Municipal de Abastecimiento y Saneamiento de Aguas de Sevilla, SA corresponde a regadíos, un 18% a abastecimiento de poblaciones e industrias y el 14% des **cisternas** que sirven para el **abastecimiento de agua**. Los **aljibes** son un caso par

Figure 13. Concordance for *abastecimiento* in the Spanish section of our corpus

term recharge (Figure 14), which points at the common absence of terminological variants in terminological databases. *Abastecimiento* does appear in the PUERTOTERM database related to the concept RECHARGE OF AN AQUIFER (Figure 15).

IATE ID	Classification	English Term	Spanish Term
1419000	3606003 - Life sciences 6831 - Building and public works	recharge projects artificial recharge projects	proyectos de alimentación proyectos de alimentación artificial
1418997	3606003 - Life sciences 6831 - Building and public works	artificial recharge deliberate artificial recharge	alimentación artificial
1418995	3606003 - Life sciences	ground-water recharge recharge of an aquifer ground-water increment intake of ground-water	recarga de un acuífero alimentación de un acuífero

Figure 14. Search results of the term *recharge* in the IATE database

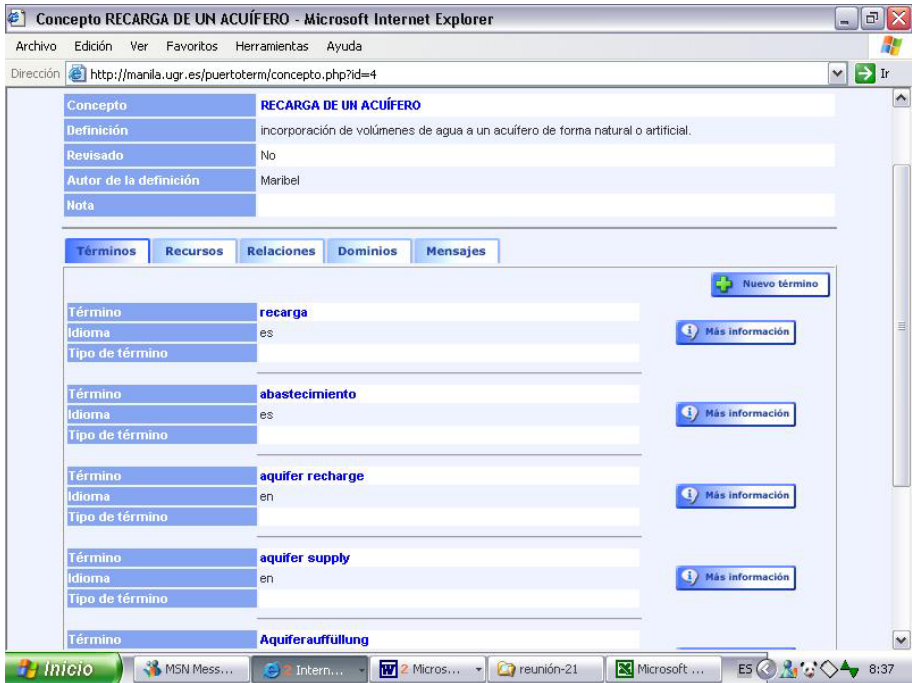


Figure 15. The concept RECHARGE OF AN AQUIFER in the PUERTOTERM knowledge base

4. Highlighting the multidimensionality of terms

Focussing on corpus data also highlights the multidimensionality of the terms within a domain (Bowker and Meyer 1993). 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; concordances show the different activations of each concept in real texts.

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, encyclopaedias (Figure 16) and databases, concordances (Figure 17) allow the identification of more parameters for classification (DIRECTION, HEIGHT, SPEED, INTENSITY, SCALE, etc.).

If we take one of those parameters, for instance INTENSITY OF WIND, 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

The screenshot shows the Enciclonet website interface. At the top, there are logos for 'enciclonet', 'Auto Escuela multimedia', and 'Cómprala en...'. Below the navigation bar, the page title is 'viento'. The left sidebar contains a 'consulta' section with instructions, a 'registro de usuario' section with the username 'unigra001' and email 'sus: 27/5/2006', and a 'finalizar sesión' button. The main content area displays the definition of 'viento' and a list of wind types: Alisios, Westerlies, Vientos periódicos, Monzón, Brisa, Vientos locales y regionales (Föhn, Bora, Mistral, Etesio, Vardarac, Burán, Pampero, Khamsin o Chamsin, Siroco, Brickfielder, Sonora, Suestado, Elefanta), and 'Distribución planetaria de los grandes vientos' and 'Influencias del viento sobre el relieve terrestre'. A right sidebar contains a 'siguiente' section with a definition of 'viento' and a 'viento' section with a definition of 'viento'.

Figure 16. Entry for *viento* in *Enciclonet*

name these grades: *calma*, *ventolina*, *flojito*, *flojo*, *bonancible*, *fresquito*, *fresco*, [...] *temporal muy duro* and *temporal huracanado*. The 324 hits for *viento* provided by the IATE database include expressions such as *viento de mar*, *mar de viento*, *frescachón*, *viento fuerte*, *viento de cara*, etc. Nevertheless, neither encyclopaedias nor databases offer a clue about classification parameters, essential to terminology management.

However, if we take a look at concordances, we can identify more types of wind (*viento*), and the elements phraseologically relevant for the keyword *viento*, including: (a) adjectives (*calmados*, *débiles*, *intensos*, *fuertes*, *moderados*) and (b) prepositional phrases (*de intensidades crecientes*). 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 17, we can see the concordances of the search item *viento** (wind) tagged for the classification parameters of: DIRECTION, HEIGHT, SPEED, INTENSITY, CONVERGENCE, POSITIVE/NEGATIVE EFFECTS, SCALE, PLACE, HUMIDITY/TEMPERATURE, FREQUENCY, PREVALENCE (Tercedor Sánchez and López Rodríguez 2006).

bien entrado el siglo XIX que eran vientos de procedencia sahariana empezaron a aparecer sectores con vientos de dirección Este en los ni	Direction
s que los circundan, es decir, son vientos ascendentes. Aparece sobre mente. En el caso de un valle, los vientos descendentes se pueden proa	Height
ió seguramente de la incidencia de vientos de altas velocidades y de u de materiales capaces de resistir vientos de hasta 240 k.p.h. Dichos	Speed
a mañana, con el cielo despejado y vientos calmados, cuando el efecto área casi libre de nubosidad con vientos débiles en un radio de acci localizan las zonas de calmas, con vientos flojos aunque con actividad mayores daños son causados por los vientos fuertes, lluvias intensas Chorro (jet) polar Cinturón de vientos intensos y, preferentemente rsión de temperatura) produjeron vientos ligeros y nieblas densas. E	Intensity
mantenimiento normal. – Resistir vientos moderados. – Poder servir, DEL OESTE: Cinturones amplios de vientos persistentes con un compone S) y que favorece la aparición de vientos suaves y de tormentas, con , sobre una duna incipiente actúan vientos de intensidades crecientes n “grado 12” correspondiente a los vientos de temporal huracanado dond	Convergence
a, es decir, en zonas donde se dan vientos convergentes. Los vientos rea de presión relativa máxima con vientos divergentes rotando en sent ermedias. . En tanques sujetos a vientos benignos, corrientes de den gicos, atenúa el efecto microbi so vientos peligrosos cuando alcanzan	Possitive-negative effects
en superficie. Cuando se trata de vientos planetarios los mecanismos los siguientes: El régimen de los vientos locales, reinantes y domina iones. 6. BRISAS TÉRMICAS: Son vientos costeros debidos a la difer as aproximadamente paralelas a los vientos pamperos. Estas acumulación	Scale
amsin o Chamsin Entramos ya en los vientos cálidos y secos. Procedente ropopausa, pasada la región de los vientos helados, se encuentra la	Place
udeste en el hemisferio sur. Estos vientos constantes se llaman viento CFB S Figura 3.26b. Rosas de los vientos multianuales en San José de	Temperature-humidity
geramente oblicua respecto a los vientos dominantes de componente W	Frequency
mo de 7,2 mm/día en enero. Los vientos prevalectientes soplan desde	Prevalence

Figure 17. Filtered concordance displaying the multidimensionality of the term *viento* (wind)

5. Conclusions

In this paper we have focused on the methodology of the PUERTOTERM project and on the possibilities of concordance analysis for terminographical work. Concordances may be used to acquire expert knowledge and to understand the relevant concepts 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.

Following a bottom-up, 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.

Acknowledgements

This research has been carried out within the framework of the projects (1) *MARCO COSTA: Marcos de conocimiento multilingüe en la gestión integrada de zonas costeras (P06-HUM-01481)*, funded by the Andalusian Regional government, with a duration of three years from 2007, and (2) *PUERTOTERM: Knowledge representation and the generation of terminological resources in the domain of Coastal Engineering (BFF 2003-04720)*. This paper is an extended version of Tercedor and López (2006).

Notes

1. PUERTOTERM: Knowledge representation and the generation of terminological resources in the domain of Coastal Engineering (BFF 2003-04720) is a R&D project funded by the Spanish Ministry of Education. It has been developed between 2003 and 2006. This project has evolved into a new research project called *MarcoCosta: Marcos de conocimiento multilingüe en la gestión integrada de zonas costeras*.
2. The knowledge base is still at the testing stage and is only accessible to researchers.
3. Also known as Lexical Grammar of Martín Mingorance.
4. WordSmith Tools: <http://www.lexically.net>
5. The frame notion has been further developed in Faber et al. (2006), where frames are defined as a type of cognitive structuring device based on experience that provide the background knowledge and motivation for the existence of words in a language as well as the way those words are used in discourse.
6. Andalusian Center for Coastal Studies: <http://www.puertosycostas.com>
7. Agent is not to be understood under a Case Grammar perspective. An agent is an animate or inanimate entity that is the doer of an event.
8. We have proposed a closed list of relations: AFFECTS, ATTRIBUTE OF, MADE OF, OPPOSITE TO, DELIMITED BY, STUDIES, MEASURES, PART OF, REPRESENTS, RESULT OF, IS CARRIED OUT WITH, HAS FUNCTION, TAKES PLACE IN, TYPE OF, IS LOCATED IN.
9. This notion follows previous research in terminology by Meyer and Mackintosh (1996), who propose the use of knowledge probes to extract superordinate terms, and Pearson's (1998) formulation of terminographic definitions based on such expressions.
10. WordSmith Tools retrieves morphological variants having the same stem through adding a wildcard (*) after the stem. Therefore irregular forms are not taken into account.

References

- Atkins, B.T.S., C.J. Fillmore and C.R. Johnson. 2003. "Lexicographic relevance: selecting information from corpus evidence." *International Journal of Lexicography* 16(3), 251–280.
- Barsalou, L.W. 2003. "Situating simulation in the human conceptual system." *Language and cognitive processes* 18(5/6), 513–562.
- Bowler, L. and I. Meyer. 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*. 123–137. Frankfurt: Indeks Verlag.
- Faber, P., C. López Rodríguez and M.I. Tercedor Sánchez. 2001. "Utilización de técnicas de corpus en la representación del conocimiento médico." *Terminology* 7(2), 167–197.
- Faber, P. and R. Mairal Usón. 1999. *Constructing a Lexicon of English Verbs*. Berlin: Mouton de Gruyter.
- Faber, P., C. Márquez and M. Vega. 2005. "Framing terminology: A process-oriented approach." *Meta* 50(4). [<http://www.erudit.org/livre/meta/2005/000255co.pdf>]. Accessed November 2006.
- Faber, P., S. Montero Martínez, M. R. Castro Prieto, J. Senso Ruiz, J.A. Prieto Velasco, P. León Aráuz, C. Márquez Linares, M. Vega Expósito. 2006. "Process-oriented terminology management in the domain of Coastal Engineering." In L'Homme, M.C. (ed.). *Processing of Terms in Specialized Dictionaries: New models and techniques* (Special issue of *Terminology*) 12(2), 189–213.
- 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* 280(1), 20–32. New York: NY Academy of Sciences.
- Fillmore, C. J. 1982. "Frame semantics" In Linguistic Society of Korea (ed.). *Linguistics in the Morning Calm*. 111–137. Seoul: Hanshin.
- Fillmore, C.J., C.R. Johnson and M. Petruck. 2003. "Background to FrameNet." *International Journal of Lexicography* 16(3), 235–250.
- IATE. *Terminological Data Bank of the European Institutions*. <https://iate.cdt.eu.int/iatenew/consultation/search/sresults.jsp?PAGE=1>. Accessed November 2006.
- López Rodríguez, C.I. 2002. "Training translators to learn from news report corpora: the case of Anglo-American cultural references." In Maia, B., J. Haller and M. Ulrych (eds.). *Training the Language Services Provider for the New Millennium*. 213–222. Porto: Faculdade de Letras, Universidade do Porto.
- López Rodríguez, C.I. 2003. "Electronic resources and lexical cohesion in the construction of intercultural competence." *Lebende Sprachen* 4, 152–156.
- López Rodríguez, C.I. 2007. "Understanding scientific communication through the extraction of the conceptual and rhetorical information codified by verbs." *Terminology* 13(1), 61–84.
- Louw, B. 1993. "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies." In Baker, M., G. Francis and E. Tognini-Bonelli (eds.). *Text and Technology: In honour of John Sinclair*. 157–176. Amsterdam: John Benjamins.
- Martín Mingorance, L. 1984. "Lexical fields and stepwise lexical decomposition in a contrastive English-Spanish verb valency dictionary." In Hartmann, R. (ed.). *LEXeter 83: Proceedings of the International Conference on Lexicography*. 226–236. Tübingen: Niemeyer.

- Martín Mingorance, L. 1989. "Functional Grammar and Lexematics." In Tomaszczyk, J. and B. Lewandowska (eds.). *Meaning and Lexicography*. 227–253. Amsterdam / Philadelphia: John Benjamins.
- Martín Mingorance, L. 1995. "Lexical logic and structural semantics: Methodological underpinnings in the structuring of a lexical database for natural language processing." Hoinkes, U. (ed.). *Panorama der Lexikalischen Semantik*. 461–474. Tübinga: Gunter Narr.
- Meiro, G. 2001–2005. *Enciclonet*. <http://www.enciclonet.com>. Accessed November 2006.
- Meyer, I. and K. Macintosh. 1996. "Refining the terminographer's concept-analysis methods: How can phraseology help?" *Terminology* 3(1), 1–26.
- Pearson, J. 1998. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.
- Moreno Ortiz, A. and C. Pérez Hernández. 2000. "Reusing the Mikrokosmos ontology for concept-based multilingual terminology databases." In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. 1061–1067. Athens, Greece.
- Pérez Hernández, M.C. 2002. *Explotación de los corpora textuales informatizados para la creación de bases terminológicas basadas en el conocimiento*, vol. 18. Madrid: CSIC / Elies. <http://elies.rediris.es/elies18>. Accessed December 2006.
- Prieto Velasco, J.A. 2006. *Información gráfica y grados de especialidad en el discurso científico-técnico: un estudio de corpus*. Unpublished MA dissertation presented at the University of Granada.
- Prieto Velasco, J.A. In press. "Improving scientific and technical translations through illustrations." In *Proceedings of the VIIIth Leipziger Internationale Konferenz zu Grundfragen der Translatologie „Translationsqualität“. Lebende Sprachen*.
- Prieto Velasco, J. A. 2007. "Visualizar para traducir: cómo gestionar la terminología en traducción científico-técnica." In Balbuena Tomezano, M.C. and A. García Calderón (eds.). *Traducción y mediación cultural: reflexiones interdisciplinares*. 357–368. Granada: Atrio.
- Temmerman, R. and K. Kerremans. 2003. "Terminography: Ontology building and the sociocognitive approach to terminology description." In *XVIIe Congrès International des Linguistes (CIL17-conference)*. Prague. http://www.hf.uib.no/forskingskole/temmerman_art-prague03.pdf. Accessed October 2006.
- Tercedor Sánchez, M.I. and C.I. López Rodríguez. 2006. "Retrieving and codifying lexical information in process oriented terminology management." In Corino, E., C. Marelllo and C. Onesti (eds.). *Proceedings XII Euralex International Congress*. 837–845. Alessandria: Edizioni dell'Orso.

Authors' addresses

Maribel Tercedor
 Facultad de Traducción e Interpretación
 Departamento de Traducción e
 Interpretación. Buensuceso 11. Granada
 18071, España.
 Universidad de Granada.
 itercedo@ugr.es

Clara I. López-Rodríguez
 Facultad de Traducción e Interpretación
 Departamento de Traducción e
 Interpretación. Buensuceso 11. Granada
 18071, España.
 Universidad de Granada.
 clarailr@ugr.es

About the authors

Maribel Tercedor is a full professor in Translation at the University of Granada, Spain, where she teaches Scientific, Technical and Audiovisual Translation. Her main research deals with codification of information in knowledge bases and Scientific and Technical translation. She is also carrying out research into lexical aspects of describing information from images.

Clara I. López Rodríguez teaches scientific and technical translation at the Faculty of Translation and Interpreting of the University of Granada (Spain), where she is a senior professor. She holds degrees from the University of Granada and Portsmouth. Her PhD thesis dealt with the relation between lexical cohesion, text type and medical translation. Her current research deals with scientific translation, and the application of corpus linguistics to terminology and translation.