LexTec – a rich language resource for technical domains in Portuguese

Palmira Marrafa, Raquel Amaro, Sara Mendes

Group for the Computation of Lexical and Grammatical Knowledge
Centro de Linguística da Universidade de Lisboa
Avenida Professor Gama Pinto, 2, 1649-003 Lisboa - Portugal
palmira.marrafa@netcabo.pt, {ramaro, sara.mendes}@clul.ul.pt

Abstract

The growing amount of available information and the importance given to the access to technical information enhance the potential role of NLP applications in enabling users to deal with information for a variety of knowledge domains. In this process, language resources are crucial. This paper presents Lextec, a rich computational language resource for technical vocabulary in Portuguese. Encoding a representative set of terms for ten different technical domains, this concept-based relational language resource combines a wide range of linguistic information by integrating each entry in a domain-specific wordnet and associating it with a precise definition for each lexicalization in the technical domain at stake, illustrative texts and information for translation into English.

Keywords: technical lexica, relational models of the lexicon, concept-based language resources

1. Motivation

With a strong psychological motivation, relational models of the lexicon such as WordNet (Miller *et al.*, 1990; Fellbaum, 1998) have played a leading role in machine lexical knowledge representation in the last decades. WordNet potential as a resource for Natural Language Processing (NLP) has also been explored in tasks associated to domain-specific information, such as systems for information extraction and document indexing, retrieval and preservation, and applications for technical domains such as Law (Peters *et al.*, 2006), Medicine (Elhadad & Sutaria, 2007) or Urbanism (Lacasta *et al.*, 2008). In the context of such work, WordNet shortcomings when it comes to the representation of domain-specific lexical units have been pointed out by several authors (Bodenreider *et al.*, 2003; Smith & Fellbaum, 2004, among others). Among the most salient, we underline the lack of domain expertise of the lexicographers developing it as mirrored in the quality of domain-specific lexicalizations represented in WordNet, which is a shortcoming totally unrelated to the model itself, and a consequence of the fact that WordNet was not originally built for domain-specific applications (Smith & Fellbaum, 2004).

Moreover, the potential of the WordNet model to represent technical concepts is made apparent by research showing that concept-based language resources (ontologies, thesauri or wordnets) have great usability in teaching or in improving the understanding of specialized contents (see Mudraya (2006) and Fuentes (2001), among many others).

In the last years, this has led to the acknowledgement of the importance of encoding domain-specific information in concept-based relational resources such as wordnets, and thus to different research efforts, such as the integration of domain-specific information into generic synsets in previously existing wordnets (Vossen, 2001; Magnini & Cavaglià, 2000) or the development of dedicated wordnets for technical domains (e.g. Smith & Fellbaum, 2004; Giunchiglia *et al.*, 2009; Roventini & Marinelli, 2004, to name a few).

The resource presented in this paper is also framed by this research effort. We will show how LexTec provides Portuguese with rich specialized lexica for ten technical domains, thus contributing to softening the lexical bottleneck for Portuguese in the area of domain-specific language resources, while contributing to make apparent the suitability of the WordNet model for representing domain-specific lexical information, i.e. terminology.

2. The LexTec database

LexTec can be generally defined as a language resource combining a set of domain-specific wordnets for European Portuguese, providing additional information regarding the technical lexical units covered.

Following from research on the extension of WordNet.PT (Marrafa 2001, 2002) to technical domains, the LexTec database covers 10 technical domains – Banking, Commerce, Construction, Economy and Business Management, Energy, Environment, Insurance, International Trade Law, Telecommunications, and Tourism. Each domain-specific wordnet has been independently built, opening up the possibility of augmenting the coverage of the resource to other domains, as has happened in the past: the project initially covered 4 domains, was then augmented to 8 domains and currently covers 10.

For each domain, a glossary of technical vocabulary, i.e. an ordered list of domain-specific lexical units, is available to the user. Each lexical expression represented in the resource is integrated in the relevant domain-specific wordnet, and thus explicitly related to the lexicalizations of other domain-specific concepts also included in this language resource, as well as with general lexicon synsets inherited from WordNet.PT (see Section 2.2 for more details), besides being associated to a precise definition of the concept represented in the technical domain at stake. Usage information is also provided via a system of registry tags, which mark expressions with

information regarding, for instance, their origin and pragmatic context of use. Additional usage information is also provided by the association of each synset variant to texts illustrating the contexts in which the expression at stake is used, sometimes also providing additional information contributing to a more precise grasp of the concept represented. Finally, all expressions are linked to their equivalent in English, increasing the potential uses of this resource, both for human translation purposes and for its integration in multilingual applications.

2.1 Data selection

Each of the 10 technical domains represented in LexTec includes an average of 1000 lexical expressions. As a whole, the resource involves the codification of information regarding more than 10 000 domain-specifc lexical units, and includes the specification of over 20 000 lexical-conceptual relations.

Terms to be included in the resource were selected based on the analysis of domain-specific *corpora* data. Given the unavailability of domain-specific *corpora* for Portuguese for the technical domains represented in LexTec, for the construction of the LexTec database we used *corpora* composed of texts collected from the Web, covering scientific and academic dissemination documents, such as papers or thesis; professional dissemination and advertising texts; domain-specific journals and other publications. Candidate terms were extracted from each domain-specific *corpora* according to two main criteria: first, the semantic-domain approach was followed, i.e. terms related to technical vocabulary already selected for integration in the resource were preferred; and secondly, frequency of occurrence of candidate terms was also taken into account, as we assume the most representative vocabulary for a given technical domain to be more frequently used.

While the second main criterion mentioned above for the selection of candidate terms essentially self-justifies itself, the same is not necessarily true for the first one. The use of the semantic-domain approach in the construction of language resources, and specifically of concept-based relational resources such as the one presented in this paper, typically aims at avoiding flat wordnets, i.e. a relational language resource mainly composed of independent and unlinked nodes. This is particularly undesireable in the case of this type of language resource since one of the main assumptions of this kind of model of lexical representation, and one of its distinctive characteristics with regard to other language resources for technical domains such as the traditional term banks, is that the meaning of each concept covered by the system is defined by the network of relations it has with other nodes in the database. Thus, the denser the network of relations encoded in the database, the more accurate is the characterization of each concept represented.

Once the lists of candidate terms to be included in the resource were compiled according to the aforementioned criteria, the final lists of expressions to be encoded in the resource were manually selected by lexicographers and validated by specialists of each technical domain when relevant.

The *corpus* of texts compiled for the selection of candidate terms was also used as a base for defining and validating the lexical-conceptual network of relations established, as well as as a source for selecting representative examples to illustrate the use of the expressions included in the resource.

The methodology used for the definition of the network of relations for the domain-specific wordnets included in LexTec, as well as the technical details regarding the integration of all the information provided in a single resource are addressed below.

2.2 General Approach and Data implementation

Being a concept-based relational language resource, in the sense that the nodes in the network represent concepts, lexicalized by sets of synonymous lexical units, i.e. synsets, which are used as a label for these nodes, and, as briefly discussed in the previous section, whose meaning is defined by the network of relations holding between them (see Fellbaum (1998)), in the implementation stage of LexTec, decisions had to be made not only with regard to the list of terms to be included in the resource (see Section 2.1), but also in what concerns determining which expressions represented the same concept and thus should be encoded in the same synset. Additionally, the network of relations holding between synsets also had to be defined, a task we address further below.

In fact, although technical vocabulary would be expected to have a lower ratio of synonymy relations due to the precision characterizing specialized discourse, in which the "form and content of terms tends towards an unambiguous relationship" (Cabré, 1998: 116), the existence of synonymy in terminology has long been acknowledged (Daille et al., 1996; Freixa, 2002; Cabré, 2008), and is once more demonstrated by the results of the work described in this paper and, specifically, by the amount of synonymy relations identified in the aforementioned *corpora* and thus encoded in LexTec (see the average of terms per synset presented in Table 1), which show an overall average of variants per synset of 1,71, in contrast with the lower average observed for WordNet.PT, a general lexicon wordnet whose synsets have an average of 1,27 variants.

One of the factors contributing to the high ratio of synonymy relations in technical lexica is the integration of terms in foreign languages, typically English, in the terminology of other languages. Moreover, these foreign expressions typically co-exist with variants in the target language. Contrastive studies on common and technical lexica (Amaro & Mendes, 2012) have shown that synonymy is indeed a distinctive feature of technical lexica, although specificities are observed when individual domains are considered (see Table 1 for a comparison between different representative technical domains included in LexTec).

For defining the network of relations for domain-specific wordnets, in LexTec we used the

well-established and tested criteria used in the development of WordNet.PT (Marrafa, 2001; Marrafa *et al.*, 2005; Amaro *et al.*, 2010). This way, LexTec includes all the lexical-conceptual relations used in WordNet.PT: built in the EuroWordNet framework, this wordnet covers all the relations considered in the EuroWordNet model (Vossen, 1998) plus some additional cross-POS relations defined in the context of research on the organization of adjectives and verbs in the mental lexicon (see Amaro et al. (2006) and Amaro *et al.* (2010) for details on the specific contributions of the research developed under the scope of the development of WordNet.PT to the improvement of relational models of the lexicon).

		N	V	Adj.	PN	Total
Environment	lexical entries (%)	66,5%	3,0%	7,0%	23,6%	
	synsets (%)	67,5%	4,9%	11,0%	16,6%	
	average variant/synset	1,75	1,07	1,13	2,53	1,78
Energy	lexical entries (%)	78,8%	2,5%	3,8%	14,8%	
	synsets (%)	80,2%	4,3%	6,2%	9,3%	
	average variant/synset	1,77	1,01	1,10	2,87	1,80
Telecom	lexical entries (%)	77,9%	3,8%	0,9%	17,4%	
	synsets (%)	82,1%	3,3%	1,5%	13,1%	
	average variant/synset	1,98	2,38	2,76	2,76	2,08
Banking	lexical entries (%)	87,5%	2,0%	1,1%	9,4%	
	synsets (%)	87,0%	3,8%	2,3%	7,0%	
	average variant/synset	2,19	1,12	1,10	2,94	2,17
Construction	lexical entries (%)	83,2%	5,1%	5,2%	6,5%	
	synsets (%)	83,5%	6,8%	6,5%	3,2%	
	average variant/synset	1,49	1,12	1,20	3,00	1,49
Tourism	lexical entries (%)	48,1%	4,5%	4,1%	43,4%	
	synsets (%)	50,9%	5,5%	5,2%	38,3%	
	average variant/synset	1,34	1,15	1,11	1,61	1,42

Table 1: PoS distribution and synonymy relation density in LexTec per technical domain

Lexico-conceptual relations	**% in the LexTec database**
hyperonymy	24,8%
hyponymy	17,9%
meronymy	6,1%
holonynmy	5,6%
subevent	0,9%
instantiation	8,0%
characterizes/is characterizable	0,6%
relates to	1,7%
event structure	8,8%
event structure (correlations)	18,7%
characteristic	3,9%
antonymy	0,6%
near antonymy	0,7%
near synonymy	0,1%
xpos synonymy	1,6%
Density	**3,2**

Table 2: Types of lexico-conceptual relations encoded in LexTec and their distribution in the network of relations of this language resource

Regarding the lexico-conceptual relations enconded in LexTec, these cover structuring relations such as hyponymy and meronymy, but also relations regarding the structure of events (cause, subevent) and prototypical participants involved in them (agent, patient, instrument, etc.). For a list of the main types of lexico-conceptual relations encoded in the LexTec database and their representativity in the network of relations of this language resource see Table 2.

As explained above, in the process of defining and implementing LexTec domain-specific wordnets, the texts collected on the Web to select the terms to be included in this language resource were used as an unstructured knowledge base for identifying the lexico-conceptual relations listed in Table 2 using indicative cues, i.e. recurring patterns of co-occurrence in language data expressing the lexico-conceptual relations considered in this resource (see Amaro (2014) for a discussion on the use of this type of co-occurrence patterns for the extraction of lexico-conceptual relations from language data). These patterns were searched for in the *corpus* and used for defining and validating the relations encoded in the different domain-specific wordnets in LexTec.

This relational information was encoded with Synsetter, an in-house flexible wordnet development tool built to allow for the full implementation of novel research results in WordNet.PT. This computational tool for the encoding of relational databases has been

developed to straightforwardly allow for updates and improvements, such as the extension of the WordNet.PT model to the representation of technical lexicalizations. It allowed us, in particular, to build parallel independent domain-specific wordnets, anchored in high-level pre-existing general lexicon synsets.

In order to maintain the uniformity of the technical wordnets built and of WordNet.PT, additional information included in LexTec, such as the illustrative texts and the English translation of the technical terms, has been encoded in a separate, though integrated, database in XML format. This option was motivated by the importance of leaving open the possibility of a straightforward integration of the aforementioned two resources in the future, i.e. for the merging of common and technical lexicon wordnets, a task that is far from trivial, raising several challenges, but whose accomplishment would create a very useful language resource, besides having the potential of providing further insights on the organization of the mental lexicon, specifically on the integration and interactions between general and domain-specific lexical units (for a discussion on the advantages of doing so and on its feasibility see Amaro & Mendes (2012)).

For distribution and consultation, all the information included in LexTec was combined in a single SQL database, using an in-house tool that takes Synsetter and XML databases as input. This SQL database is used as the source of information for the queries launched through the web interface described in the last section of this paper.

	Nouns	**Verbs**	**Adjectives**	**Proper names**	**Total**
number of terms	6 296	291	271	1 305	8 163
number of concepts (synsets)	3 691	246	237	591	4 765
average of terms per synset	1.71	1.18	1.14	2.21	1.71
hyperonymy	6 259	270	12		6 541
meronymy	1 505			280	1 785
relations regarding event structure	3 947	400			4 347
instantiation	445			775	1 220
equivalence relations	235	162	56		453
other relations	583	4	377		964
total number of relations	2 974	836	446	1 055	15 310
average of relations per synset	3.52	3.40	1.88	1.79	3.21

Table 3: Distribution of synsets, variants and lexical-conceptual relations in LexTec

2.1 Results

The LexTec database publicly available at this stage covers more than 8000 domain-specific lexical units from the main PoS from 10 different technical domains. LexTec is balanced between the different domains represented, although full distribution of the resource is still ongoing[1]. In Table 3 we present the aggregated distribution of domain-specific synsets and variants between different PoS, as well as the distribution of some of the main lexical-conceptual relations encoded in the database.

PoS distribution in LexTec reflects what is generally assumed to be characteristic of technical domains, specifically that the description of a given domain is mainly constituted by nominal expressions (Cabré, 1998: 36): in terms of PoS distribution, we observe a larger percentage of nominal nodes in LexTec (77.5% of nouns and 12.4% of proper names), and a consequent smaller percentage of the other PoS, although the proportion between nouns and proper names, as well as other PoS, can be considerably varying in different technical domains, as can be observed in Table 1, where the type of data presented in Table 3 for the LexTec database as a whole is discriminated for a set of representative technical domains showing contrastive properties with regard to aspects such as PoS distribution. The inclusion of proper names in LexTec, a subtype of the nominal PoS often not included in language resources, is motivated by its representativity and significance in the specific technical domains (and *corpora*) analyzed.

Proper names lexicalize a wide variety of entities, from individuals, institutions, brands or companies, works of art, books or documents. The approach followed in this project made apparent the significance of some proper names, such as lexicalizations of specific laws, treaties, authorities or institutions, in certain technical domains.

The presence of proper names is not expected to affect the general usability of a language resource such as Lextec. On the contrary, besides being related in the database to common nouns through a specific relation – *instantiation* – which distinguishes the instances of this class from other nominal nodes, thus mirroring the specific and distinctive characteristics of this subtype of the nominal PoS, the inclusion of propers names in the LexTec database can potentiate its use as a language resource for a wider range of NLP applications, namely named entity recognition applications or systems performing inference tasks.

[1] Being incrementally made available to the public, the full distribution of three of the domain-specific lexica is still ongoing: Economy and Business Management; Insurance; and International Trade Law.

Finally, and besides the unbalance in terms of PoS distribution observed in the domain-specific language resource discussed in this paper, contrasts regarding the weight of different lexico-conceptual relations holding between concepts encoded in the database are also observed. Obviously, contrasts in PoS distribution and in the amount of instances of certain lexical-conceptual relations are not completely independent facts, as certain types of relations impose restrictions on the PoS of the lexical expressions they link.

In this context, we underline the weight of hyperonymy relations in the overall number of relations observed, although this is far from being unexpected in technical lexica, since the specification of concepts, expressed in wordnets through hyperonymy relations, is known to be quite productive in terminology (Daille *et al.*, 1996; Freixa, 2002; Cabré, 2008; among others).

3. Navigating Lextec online

In order to make LexTec data publicly available and freely usable, this domain-specific language resource has been released on the WWW[2] via a web interface for online consultation allowing users to navigate and easily access the data on the different technical lexica included in this resource.

The web interface for navigating LexTec database has been designed to allow for the visualization of all the information encoded. Working on top of an SQL database, this web interface allows for searching information per domain as well as per string (i.e. per term), as shown in Figures 1 and 2. It also allows for visualizing all the terms in a domain as a list (see Figure 2).

Figure 1: Homepage of the LexTec web interface, in which the user can selected for one of the technical domains available or opt for a free term search in the entire database

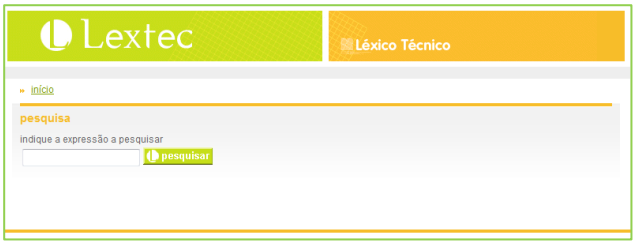

Figure 2: Information search per string in the LexTec web interface

Regarding the information available for each individual entry, the user can alternate between different windows to access the different types of information available (definition, lexical-conceptual relations, and illustrative texts – see Figure 3) and navigate in the network of relations by going from one entry to another following a given lexical-conceptual relation.

Figure 2: List of terms in the domain of *Environment* begining with the letter C represented in LexTec

4. Final Remarks

This paper details the richness and coverage of the information encoded in a computational concept-based relational language resource for technical domains in European Portuguese: LexTec. Developed following linguistically motivated and solid criteria, as detailed in the paper, this publicly available resource provides Portuguese with rich domain-specific language resources for ten technical domains, this way crucially contributing to reducing the lexical bottleneck for this language in the area of technical applications.

[2] http://www.instituto-camoes.pt/lextec

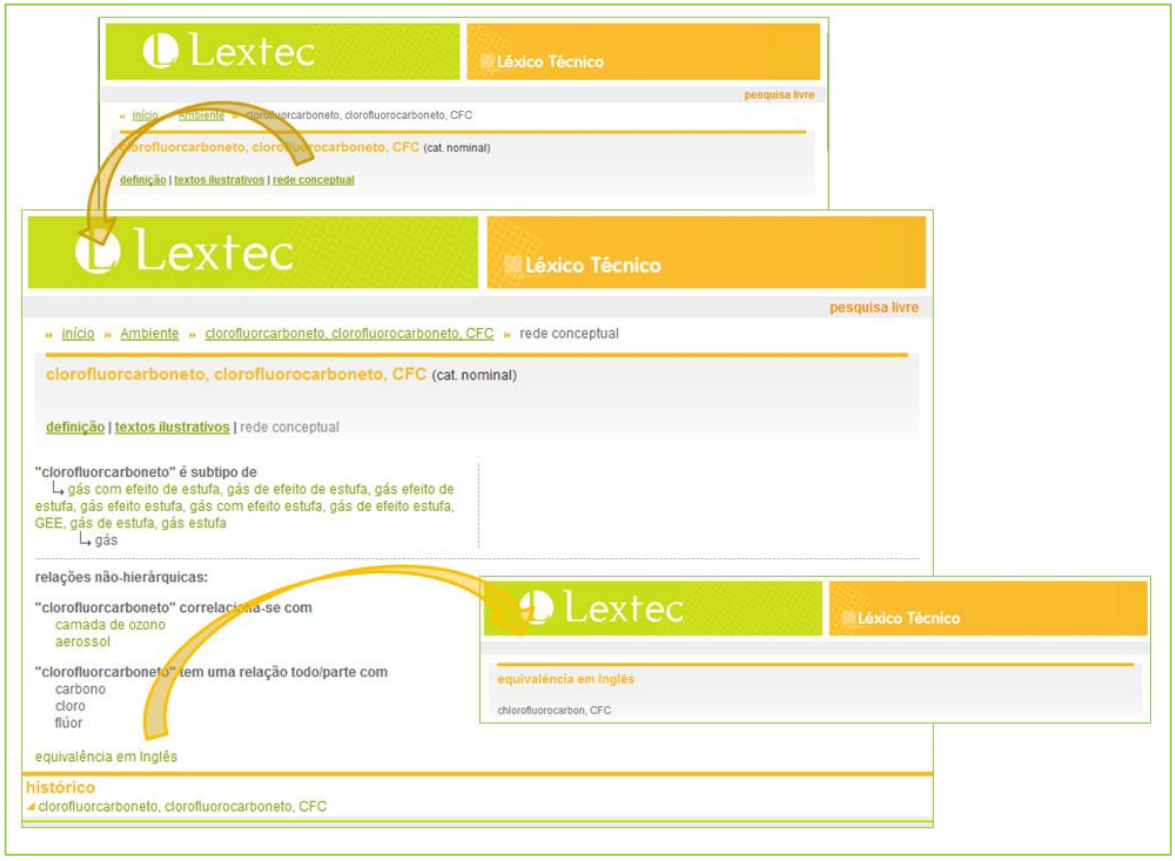

Figure 3: Overview of the different windows available in the web interface for visualizing the various types of information encoded in LexTec – definition, illustrative texts, lexical-conceptual relations, and information for translation in English –regarding specific lexical entries

5. References

Amaro, R., Chaves, R. P., Marrafa, P. & Mendes, S. (2006). Enriching wordnets with new relations and with event and argument structures. In *Proceedings of the 6th International conference on Intelligent Text Processing and Computational Linguistics CICLing-2006*, pp. 28-40.

Amaro, R. & Mendes, S. (2012). Towards merging common and technical lexicon wordnets. In *Proceedings of the 3rd Workshop on Congnitive Aspects of the Lexicon (CogALex-III) at the 24th International Conference on Computational Linguistics - COLING 2012*. Mumbai, India, pp. 147-160.

Amaro, R., Mendes, S. & Marrafa, P. (2010). Encoding Event and Argument Structures in Wordnets. In Sojka, P., Horák, A., Kopeček, I. & Pala, K. (eds.) *TSD 2010, LNAI 6231*, Berlin Heidelberg: Springer-Verlag, pp. 21-28

Bodenreider, O., Burgun, A. & Mitchell, J. A. (2003). Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. In *Studies in Health Technology and Informatics*, 95, pp. 379-384.

Buitelaar, P. & Sacaleanu, B. (2002). Extending synsets with medical terms. In *Proceedings of First Global WordNet Conference*. Mysore, India.

Cabré, M. T. (2008). El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología. In *IBÉRICA* 16, pp. 9-36.

Cabré, M. T. (1998). *Terminology. Theory, methods and applications*. Amsterdam: John Benjamins Publishing.

Daille, B., Habert, B., Jacquemin, C. & Royauté, J. (1996). Empirical observation of term variations and principles for their description. In *Terminology* 3 (2), pp. 197-257.

Fellbaum, C. (1998) (ed.) *WordNet: an electronical lexical database*. Cambridge: The MIT Press.

Freixa, J. (2002). *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient*. PhD dissertation, Universitat Pompeu Fabra, Barcelona.

Fuentes, A. C. (2001). Lexical Behaviour in Academic and Technical Corpora: Implications for ESP Development. In *Language Learning & Tecnhology*, vol.5, nº 3, pp. 106-129.

Giunchiglia, F., Maltese, V., Farazi, F. & Dutta, B. (2009). *GeoWordNet: a resource for geo-spatial applications*. Technical Report #DISI-09-071: http://eprints.biblio.unitn.it/1777/1/071.pdf.

Lacasta, J., Nogueras-Isso, J., Zarazaga-Soria, P. & Muro-Medrano, R. (2008). Generating an urban domain ontology through the merging of cross-domain lexical ontologies. In *Conceptual Models for Urban Practitioners*. Bologna: Società Editrice Esculapio, pp. 69-84.

Magnini, B. & Cavaglià, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece.

Magnini, B. & Strapparava, C. (2001). Using WordNet to improve user modelling in a web document recommender system. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Pittsburgh, Pennsylvania.

Marrafa, P. (2002). The Portuguese WordNet: General Architecture and Semantic Internal Relations. In *DELTA*.

Marrafa, P. (2001). *WordNet do Português - Uma base de dados de conhecimento linguístico*. Instituto Camões.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C. & Mendes, S. (2005). WordNet.PT - Uma Rede Léxico-conceptual do Português on-line. Manuscript presented at XXI Encontro da Associação Portuguesa de Linguística, September 28-30, 2005, Oporto, Portugal.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K.J. (1990). Introduction to WordNet: An online lexical database. In *International Journal of Lexicography, 3*(4), pp. 235-244.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. In *English for Specific Purposes* 25. Elsevier, pp. 235-256.

Peters, W., Sagri, M., Tiscornia D. & Castagnoli, S. (2006). The LOIS Project. In *Proceedings of Linguistic Resources Evaluation Conference (*LREC'06). Genova, Italy, pp. 23-27.

Roventini, A. & Marinelli, R. (2004). Extending the Italian WordNet with the Specialized Language of the Maritime Domain. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum & P. Vossen (Eds.). *Proceedings of the Global WordNet Conference 2004 (GWC 2004)*. Brno: Masaryk University, pp. 193-198.

Smith, B. & Fellbaum, C. (2004). Medical WordNet: a New Methodology for the Construction and Validation of Information Resources for Consumer Health. In *Proceedings of COLING 2004*. Geneva, Switzerland.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, 32, 73–89. Reprinted in Vossen, P. (ed.) *EuroWordNet – A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Kluwer Academic Publishers.

Vossen, P. (2001). Extending, trimming and fusing wordnet for technical documents. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources.* Pittsburgh, Pennsylvania.