The TermiNet Project: an Overview

Ariani Di Felippo
Interinstitutional Center for Research and Development in Computational Linguistics (NILC)/
Research Group of Terminology (GETerm), Federal University of São Carlos (UFSCar)
Rodovia Washington Luís, km 235 - SP-310
CP 676, 13565-905, São Carlos, SP, Brazil
ariani@ufscar.br

Abstract

Linguistic resources with domain-specific coverage are crucial for the development of concrete Natural Language Processing (NLP) systems. In this paper we give a global introduction to the ongoing (since 2009) TermiNet project, whose aims are to instantiate a generic NLP methodology for the development of terminological wordnets and to apply the instantiated methodology for building a terminological wordnet in Brazilian Portuguese.

1 Introduction

In knowledge-based Natural Language Processing (NLP) systems, the lexical knowledge database is responsible for providing, to the processing modules, the lexical units of the language and their morphological, syntactic, semantic-conceptual and even illocutionary properties (Hanks, 2004).

In this scenario, there is an increasing need of accurate general lexical-conceptual resources for developing NLP applications.

A revolutionary development of the 1990s was the Princeton WordNet (WN.Pr) (Fellbaum, 1998), an online reference lexical database built for North-American English that combines the design of a dictionary and a *thesaurus* with a rich ontological potential.

Specifically, WN.Pr is a semantic network, in which the meanings of nouns, verbs, adjectives, and adverbs are organized into "sets of cognitive synonyms" (or synsets), each expressing a distinct concept. Synsets are interlinked through conceptual-semantic (i.e., hypernymy[1]/hyponymy[2], holonymy/meronymy, entailment[3], and cause[4]) and lexical (i.e., antonymy) relations. Moreover, WN.Pr encodes a co-text sentence for each word-form in a synset and a concept gloss for each synset (i.e., an informal lexicographic definition of the concept evoked by the synset).

The success of WN.Pr is largely due to its accessibility, linguistic adequacy and potential in terms of NLP. Given that, WN.Pr serves as a model for similarly conceived wordnets in several languages. In other words, the success of WN.Pr has determined the emergence of several projects that aim the construction of wordnets for other languages than English or to develop multilingual wordnets (the most important project in this line is EuroWordNet) (Vossen, 2002).

Many recent projects with the objective of (i) integrating generic and specialized wordnets (e.g., Magnin and Speranza, 2001; Roventini and Marinelli, 2004; Bentivogli et al., 2004), (ii) enriching generic wordnets with terminological units (e.g., Buitelaar and Sacaleanu, 2002) or (iii) constructing terminological wordnets (e.g.: Sagri et al., 2004; Smith and Fellbaum, 2004) have shown that con-

[1] The term Y is a hypernym of the term X if the entity denoted by X is a (kind of) entity denoted byY.

[2] If the term Y is a hypernym of the term X then the term X is a hyponym of Y.

[3] The action A1 denoted by the verb X entails the action A2 denoted by the verb Y if A1 cannot be done unless A2 is, or has been, done

[4] The action A1 denoted by the verb X causes the action A2 denoted by the verb Y.

Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pages 92–99, Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

crete NLP application must be able to comprehend both expert and non-expert vocabulary.

Despite the existence of a reasonable number of terminological wordnets, there is no a general methodology for building this type of lexical database. Thus, motivated by this gap and by the fact that Brazilian Potuguese (PB) is a resource-poor language, the two-years TermiNet project has been developed since September 2009.

This paper gives an overview of the TermiNet project. Accordingly, in Section 2 we brief describe the original WN.Pr design that motivated the project. In Section 3 we present the aims of the TermiNet project and its methodological approach. In Section 4 we depict the current state of the project. In Section 5 we describe future work, and in Section 6 we outline potential points for collaboration with researchers from the rest of the Americas.

2 Princeton WordNet and its Design

WN.Pr contains information about nouns, verbs, adjectives and adverbs in North-American English and is organized around the notion of a *synset*. As mentioned, a synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, {car; auto; automobile; machine; motorcar} form a synset because they can be used to refer to the same concept. A synset is often further described by a concept gloss[5], e.g.: "4-wheeled; usually propelled by an internal combustion engine".

Finally, synsets can be related to each other by the conceptual-semantic relations of hyperonymy/ hyponymy, holonymy/meronymy, entailment and cause, and the lexical relation of antonymy.

In the example, taken from WN.Pr (2.1), the synset {car; auto; automobile; machine; motorcar} is related to:

(i) more general concepts or the hyperonym synset: {motor vehicle; automotive vehicle};

(ii) more specific concepts or hyponym synsets: e.g. {cruiser; squad car; patrol car; police car; prowl car} and {cab; taxi; hack; taxicab}; and

(iii) parts it is composed of: e.g. {bumper}; {car door}, {car mirror} and {car window}.

WN.Pr also includes an English co-text sentence for each word-form in a synset, and a semantic type for each synset.

Based on WN.Pr design, Brazilian Portuguese WordNet (WordNet.Br or WN.Br) project launched in 2003 departed from a previous lexical resource: the Brazilian Portuguese Thesaurus (Dias-da-Silva et al, 2002). The original WN.Br database is currently being refined, augmented, and upgraded. The improvements include the encoding of the following bits of information in to the database: (a) the co-text sentence for each word-form in a synset; (b) the concept gloss for each synset; and (c) the relevant language-independent hierarchical conceptual-semantic relations.

The current WN.Br database presents the following figures: 11,000 verb forms (4,000 synsets), 17,000 noun forms (8,000 synsets), 15,000 adjective forms (6,000 synsets), and 1,000 adverb forms (500 synsets), amounting to 44,000 word forms and 18,500 synsets (Dias-da-Silva et al, 2008).

3 The TermiNet Project

The TermiNet ("Terminological WordNet") project started in September 2009 and shall be finished finish in August 2011. It has been developed in the laboratory of the Research Group of Terminology[6] (GETerm) in Federal University of São Carlos (UFSCar) with the collaboration of the Interinstitutional Center for Research and Development in Computational Linguistics[7] (NILC/University of São Paulo) researchers.

The TermiNet project has two main objectives. The first is to instantiate the generic NLP methodology, proposed by Dias-da-Silva (2006), for developing terminological databases according to the WN.Pr model. Such methodology distinguishes itself by conciliating the linguistic and computational facets of the NLP researches. The second is to apply the instantiated methodology to build a terminological wordnet or terminet[8] in BP, since BP is a resource-poor language in NLP for which domain-specific databases in wordnet format have not been built yet.

It is important to emphasize that the main terminological resources in BP, which are availa-

[5] An informal lexicographic definition of the concept evoked by the synset.

[6] http://www.geterm.ufscar.br/

[7] http://www.nilc.icmc.usp.br

[8] In the TermiNet project, a terminological wordnet database is called "terminet".

ble through the OntoLP[9] website, are in fact (formal) ontologies or taxonomies. There is no nological WordNet-like database in BP.

In order to achieve its objectives, TermiNet has, apart from the project leader (Prof. Ariani Di Felippo), an interdisciplinary team that includes six undergraduate students: five from Linguistics and one from Computer Science courses. The Linguistics students are responsible for specific linguistic tasks in the project, such as: (i) *corpus* compilation, (ii) candidate terms extraction, (iii) synonymy identification, and (iv) semantic-conceptual relations extraction (hypernymy/hyponymy). The responsability of the Computer Science student is to support the automatic processing related to the linguistic (e.g., *tagging, parsing,* term extraction, etc.) and linguistic-computational domains during the initial stages of the project.

Moreover, the project counts with the collaboration of four PhD researchers from NILC. Specifically, TermiNet has the support of Prof. Gladis Maria de Barcellos Almeida, a specialist in terminological research and the coordinator of GETerm; Prof. Maria da Graças Volpe Nunes, the coordinator of NILC and one of the most important Brazilian NLP researchers; Prof. Sandra Aluisio, a specialist in *corpus* construction, and Prof. Thiago Pardo, who has interests in the development of lexical resources for the automatic processing of BP.

3.1 Instantiation of the NLP Tree-Domain Methodology

Based on Expert Systems development, Dias-da-Silva (2006) established a three-domain approach methodology to develop any research in NLP domain, assuming a compromise between Human Language Technology and Linguistics (Dias-da-Silva, 1998).

The linguistic-related information to be computationally modeled is likened to a rare metal. So, it must be "mined", "molded", and "assembled" into a computer-tractable system (Durkin, 1994). Accordingly, the processes of designing and implementing a terminet lexical database have to be developed in the following complementary domains: the linguistic domain, the linguistic-computational domain, and implementational or computational domain.

(a) The Linguistic-related Domain

In this domain, the lexical resources and the lexical-conceptual knowledge are mined. More specifically, the research activities in the linguistic domain are divided in two processes: the selection of the lexical resources for building the terminet database, and the specification of the lexical-conceptual knowledge that characterize a terminet.

The linguist starts off these procedures by delimitating the specialized domain that will be encoded in wordnet format.

According to Almeida and Correia (2008), dealing with an entire specialized domain is a very problematic task because the domains (e.g.: Materials Engineering) in general are composed of subdomains (e.g.: Ceramic Materials, Polymers and Metals) with different characteristics, generating a large universe of sources from which the lexical-conceptual knowledge will have to be mined.

Consequently, the authors present some criteria that may lead to delimitate a specialized domain: (i) the interest of the domain experts by terminological products (in this case, by a terminet); (ii) the relevance of the domain in the educational, social, political, economic, scientific and/or technological scenarios, and (iii) the availability of specialized resources in digital format from which the lexical-conceptual knowledge will be extracted. After delimitating the domain, it is necessary to select the lexical resources describe in (iii). According to Rigau (1998), the two main sources of information for building wide-coverage lexicons for NLP systems are: structured resources (e.g.: conventional monolingual and bilingual dictionaries, *thesauri,* taxonomies, vocabularies, etc.) and unstructured resources (i.e., *corpora*[10]).

Due to the unavailability of reusing structured resources, the *corpora* have become the main source of lexical knowledge (Nascimento, 2003; Agbago and Barrière, 2005; Cabré et al., 2005; Almeida, 2006). The increasing use of *corpora* in terminological researches is also due to the fact that "el carácter de término no se da per se, sino en función del uso de una unidad léxica en un contexto expresivo y situacional determinado" (Cabré, 1999: 124). Thus, in the TermiNet project, the *cor-*

[9] http://www.inf.pucrs.br/~ontolp/downloads.php

[10] "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2005).

pus is considered the main lexical resource that can be used to construct a terminet.

Although there are available several specialized *corpora*, the development of a terminet of certain domains may require the compilation of a *corpus*.

Based on the assumptions of Corpus Linguistics (Aluisio and Alemida, 2007), the construction of a *corpus* must follow three steps: (i) the *corpus* projection, i.e., the specification of the *corpus* typology according to the research purposes; (ii) the compilation of the texts that will compose the *corpus*, and (iii) the pre-processing of the *corpus* (i.e., conversion, clean-up, manipulation, and annotation of the texts).

From the *corpus,* the specialized knowledge will be extracted, i.e., the terminological units (or terms), the lexical relations, and the conceptual-semantic relations[11].

As mentioned in previous sections, the lexical units are organized into four syntactic categories in WN.Pr: verbs, nouns, adjectives and adverbs. Given the relevance of nouns in the organization of any terminology (i.e., the set of all terms related to a given subject field or discipline), we decided to restrict the construction of a terminet to the category of nouns. In other words, a terminet database, in principle, will only contain information about concepts lexicalized by nouns. Additionally, it will only encode the hyperonymy/hyponymy relations, which are the most important conceptual-semantic relations between nouns. The co-text sentence for each word-form in a synset and the concept gloss for each synset will not be focused in building a terminet.

As the TermiNet a *corpus*-based project, we will apply approaches and strategies to automatically recognize and extract candidate terms and relations from *corpus*.

In order to better understand the automatic candidate terms and extraction, it can be useful to identify two mainstream approaches to the problem. In the first approach, statistical measures have been proposed to define the degree of *termhood* of candidate terms, i.e., to find appropriate measures that can help in selecting good terms from a list of candidates. In the second approach, computational terminologists have tried to define, identify and recognize terms looking at pure linguistic properties, using linguistic filtering techniques aiming to identify specific syntactic term patterns (Bernhard, 2006; Pazienza et al., 2005; Cabré et al., 2001).

Once extrated, the candidate terms have be validated. Two validation estrategies will be considered in the TermiNet project. The first strategy consists on manually validating by domain experts. The second consists on automatically comparing the list of candidate terms with a list of lexical unities extracted from a general *corpus* in BP.

The automatic acquisition of hyperonym/hyponymy relation from *corpus* is commonly based on linguistic methods. These methods look for linguistic clues that indisputably indicate the relation of interest (Hearst, 1992). The linguistic clues are basically lexico-syntactic patters such as: [NP such {NP,}*{(or|and)} NP] (e.g., "works by such authors as Herrick, and Shakespeare"). The hierarchical relations extrated from *corpus* are commonly validated by domain experts.

(b) The Linguistic-Computational Domain

In this domain, the overall information selected and organized in the preceding domain is molded into a computer-tractable representation; in the case of a WordNet-like database, the computer-tractable representation is based on the notions of:

- *word form* – a orthographic representation of an individual word or a string of individual words joined with underscore characters;
- *synset* – a set of words built on the basis of the notion of synonymy in context, i.e. word interchangeability in some context;
- *lexical matrix* – associations of sets of word forms and the concepts they lexicalize;
- *relational pointers* – formal representations of the relations between the word forms in a synset and other synsets; synonymy of word forms is implicit by inclusion in the same synset; hyperonymy always relates one synset to another, and is an example of a semantic relation; hyperonymy, in particular, is represented by reflexive pointers (i.e., if a synset contains a pointer to another synset, the other synset should contain a corresponding reflexive pointer back to the original synset).

(c) The Computational Domain

In this domain, the computer-tractable representations are assembled by utilities (i.e., a computational tool to create and edit lexical knowledge). In

[11] The glosses and co-text sentences will not be specificied in the TermiNet projet.

other words, it is generated, in this domain, the terminet database. The software tool that we will use to generate the terminet database is under investigation.

4 TermiNet: Past and Current Stages of Development

The project, which started in September 2009, is still in its early stages. Consequently, the research tasks that have been developed so far are those related to the linguistic domain. As described in Section 3.1a, there are several linguistic tasks in the TermiNet project. Two of them – the delimitation of the specialized domain and the *corpus* projection – are completed. In subsections 4.1 and 4.2, we present these finished processes and in 4.3 we focus on the current activity.

4.1 Delimitation of the specialized domain

DE is conventionally defined as "*any educational or learning process or system in which the teacher or instructor is separated geographically or in time from his or her students or educational resources*".

According to the second Brazilian Yearbook of Statistics on Open and Distance Education (Anuário Brasileiro Estatístico de Educação Aberta e a Distância[12]), in 2007 there were approximately 2,5 millions of students enrolled in accredited DE courses, from basic to graduate education, in 257 accredited institutions. The number of students in DE courses has grown 24.9% in relation to 2006. Thus, we can see the relevance of the DE modality in Brazil. Despite the relevance of the DE in the Brazilian educational (and political) scenario, there is no a lexical-conceptual representation of this domain, especially in a machine-readable format.

Consequently, the instantiated methodology will be validated by building DE.WordNet (DE.WN), a specialized wordnet of the Distance Education (or Distance Learning) domain in BP. The construction of such database has been supported by domain experts from the "Open University of Brazil" (Universidade Aberta do Brasil – UAB) project of the Federal University of São Carlos (UFSCar).

DE.WN can be integrated into the wordnet lexical database for BP, the WordNet.Br (Dias-da-Silva et al., 2008), enriching it with domain specific knowledge.

4.2 *Corpus* projection

Following the assumptions of Corpus Linguistics described in Section 3, the *corpus* of DE domain has been constructed according to the steps: (i) *corpus* projection, (ii) *corpus* compilation, and (iii) the pre-processing of the texts.

The *corpus* typology in the TermiNet project was specified based on: (i) the conception of "*corpus*", (ii) the type of lexical resource to be built, and (iii) the project decisions (Di Felippo and Souza, 2009).

The *corpus* definition or conception is commonly related to three criteria: *representativeness, balance* and *authenticity.*

According to the *representativeness* criterion, we have been compiled a representative *corpus* of the DE domain. There have been many attempts to set the size, or at least establish a minimum number of texts, from which a specialized *corpus* may be compiled. To satisfy the *representativeness* criterion, we have been constructed a medium-large *corpus,* with at least 1 million of words.

In a specialized *corpus,* it is important to gather texts from different genres (i.e. technical-scientific, scientific divulgation, instructional, informative, and technical-administrative) and media (i.e, newswire, books, periodicals, etc.). Following the *balance* and *authenticity* criteria, we have been constructed a *corpus* with a balanced number of real texts per genre.

Besides, the format of the lexical database (i.e. a terminet) determined some characteristics of the *corpus.* Specifically, the *corpus* has to be synchronic/ contemporary, since a wordnet (terminological or not) encodes synchronic lexical-conceptual knowledge. The *corpus* has only to store written texts, since wordnets are lingwares for written language processing. Finally, the *corpus* in the TermiNet project has only to store texts from a specialized domain and in one language.

Additionally, some project decisions determined other characteristics of the *corpus.* Two initial decisions in the project were: (i) to apply semi-automatic methods of lexical-conceptual knowledge extraction, and (ii) to share the resources and results of the TermiNet project with Computational Linguistics community. As a consequency of the project decision described in (i), the *corpus* will be

[12] http://www.abraead.com.br/anuario/anuario_2008.pdf

annotated with part-of-speech (PoS) information, since some automatic extraction methods require it. As a consequency of the decision presented in (ii), the *corpus* will be available and usable as widely as possible on the *web*.

Finally, we also decided that once the *corpus* has been assembled, it will not be changed until the first version of DE.WN is ready.

Based on the typology proposed by Giouli and Peperidis (2002), the Table 1 summarizes the caracteritics of the *corpus* previously described.

Modality	Written *corpus*
Text Type	Written *corpus*
Medium	Newspapers, books, journals, manuals and others
Language coverage	Specialized *corpus*
Genre/register	Technical-scientific, scientific divulgation, instructional, informative and, technical-administrative
Language variables	Monolingual *corpus*
Markup	Annotated *corpus* (PoS annotation)
Production Community	Native speakers
Open-endedness	Closed *corpus*
Historical variation	Synchronic *corpus*
Availability	Online *corpus*

Table 1. The *corpus* design.

The specialized domain and *corpus* typology were specified by the undergraduate student responsible for the *corpus* compilation under the supervision of a PhD in Linguistics (leader of the project).

4.3 *Corpus* compilation

Currently, one undergraduate student from Linguistics has been compiled the *corpus*. Specifically, the *corpus* compilation comprises two processes: (i) the selection of resources and (ii) the collect of texts from these resources.

In the TermiNet project, the web is the main source for collecting texts of DE. The choice of web reflects the fact that web has become an unprecedented and virtually inexhaustible source of authentic natural language data for researchers in linguistics.

Although there are many computational tools that assist in gathering a considerable amount of texts on the web, the selection/collection of texts has been followed a manual process, which is composed of three steps: (i) to access a webpage whose content is important for compiling the *corpus*, (ii) to search the texts on the webpage by search queries as “distance education” and “distance learning”, and (iii) to save the text files on the computer.

In the pre-processing step, the text files in a non-machine readable format (e.g. pdf) are manually converted to text format (txt), which is readable by machines. This process is important because the lexical-conceptual knowledge will be (semi)automatically extracted from the *corpus,* and the extraction tools require a *corpus* whose texts are in *txt* format.

Data corrupted by the conversion or even unnecessary to the research (e.g. references, information about filliation, etc.) are excluded during the cleaning process. After that, the metadata or external information (e.g. authorship, publication details, genre and text type, etc.) on each text are being automatically annotated and encoded in a header. In the TermiNet project, we are using the header editor available at the “Portal de Corpus” website[13].

5 Future Work

According to the three-domain methodology, future steps will involve the following tasks of the linguistic domain: candidate terms and relations extraction (and validation).

In the TermiNet project, two specific software tools constructed based on lingustic approaches will be used to extract candidate terms from the DE *corpus*: $\mathrm{E}_{X}\mathrm{ATO}_{LP}$ (Lopes et al., 2009) and OntoLP (Ribeiro Jr., 2008). Additionally, we intend to extract the terms from *corpus* using the NSP (Ngram Statistics Package) tool (Bannerjee and Pedersen, 2003), i.e., a flexible and easy-to-use software tool that supports the identification and analysis of Ngrams.

To extract the hyperonymy and hyponymy relations, we will also use the OntoLP, which is a tool, actually a plug-in, for the ontologies editor Protégé[14], a widely used editor in the scientific community and which gives support to the construction of ontologies. The process of automatic

[13] http://www.nilc.icmc.usp.br:8180/portal/

[14] http://protege.stanford.edu/

ontology construction in the OntoLP tool also englobes the identification of hierarchical relation between the terms.

The synonymy relation will be also recognized and extracted automatically from the *corpus*. However, the automatic extraction method of such lexical relation is still under investigation.

After the acquisition of all lexical-conceptual information, we will develop the tasks or processes of the linguistic-computational and computational domains.

Among the expected results of the TermiNet projet are: (i) a methodological framework for building a specific type of *lingware*, i.e. terminological wordnets; (ii) a specialized *corpus* of the DE domain; (iii) a terminological lexical database based on the WN.Pr format of the DE domain. Moreover, there is the possibility of extending the WN.Br database through the inclusion of specialized knowledge.

Besides the benefits to NLP domain, the DE.WN may also contribute to the development of standard terminographic products (e.g., glossary, dictionary, vocabulary, etc.), of the DE domain since the organization of the lexical-conceptual knowledge is an essential step in building such products.

6 Collaborative Opportunities

We consider our experience in developing a terminet in BP as the major contribution that we can offer to other researchers in Latin America. Since the resources (i.e., *corpus* and lexical database) and tools (i.e., terms and relations extractors) that we have been used are language-dependent, they cannot be used directly for Spanish and English. But, we are willing to share our expertise on (i) compiling a terminological *corpus,* (ii) automatically extracting lexical-conceptual knowledge from *corpus,* and (iii) constructing a terminet database in order to develop similar projects for Spanish and English.

We are really interested in actively taking part in joint research projects that aim to construct terminological lexical database for Spanish or English, especially in *wordnet* format.

Collaboration of researchers from the USA that were directly involved in the development of wordnet databases (terminological or not), willing to share their experience and tools, would be welcome.

We would appreciate collaboration from researchers in the USA specifically in relation to computational programs or software tools used in building WordNet-like lexical database, which are responsible for the computer-tractable representation described in 3.1(b). The current WN.Br editing tool, which was originally designed to aid the linguist in carrying out the tasks of building synsets, selecting co-text sentences from *corpora*, and writing synset concept glosses, has been modified to aid the linguistic in carrying out the task of encoding conceptual relations. However, this editor is just able to deal with the hypernymy/hyponymy relations when they are inherited from WN.Pr through a conceptual-semantic alignment strategy (Dias-da-Silva et al, 2008). So, the WN.Br editor is not the most appropriate tool to TermiNer project tasks. Consequently, contributions to develop "a kind of" Grinder[15] for TermiNet would be welcome. We would also appreciate collaboration from re-searchers in the USA in relation to methodological approaches to enriching generic wordnets with terminological units.

Acknowledgments

We thank the Brazilian National Council for Scientific and Technological Development (CNPq) (471871/2009-5), and the State of São Paulo Research Foundation (FAPESP) (2009/06262-1) for supporting the TermiNet project. We also thank the NAACL HLT Young Investigators Workshop referees, who helped make this paper better.

References

Adriana Roventini and Rita Marinelli. 2004. Extending the Italian Wordnet with the specialized language of the maritime domain. In: *Proceedings of the 2nd International Global Wordnet Conference.* Masaryk University, Brno, 193-198.

Akakpo Agbago and Caroline Barrière. 2005. Corpus construction for Terminology. In: *Proceedings of the Corpus Lingustics Conference.* Birmingham, 14-17.

Bento Carlos Dias-da-Silva. 2006. Bridging the gap between linguistic theory and natural language processing. In: *Proceedings of the 16th International*

[15] This is the most important program used in building WN.Pr. Lexicographers make their additions and changes in the lexical source files, and the Grinder takes those files and converts them into a lexical database (in *wordnet* format).

Congress of Linguistics, 1997. Oxford: Elsevier Sciences, 1998, 1-10.

Bento Carlos Dias-da-Silva. 2006. O estudo linguístico-computacional da linguagem. *Letras de Hoje*, 41(2): 103-138.

Bento Carlos Dias-da-Silva, Ariani Di Felippo and Maria G. V. Nunes. 2008. The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In: *Proceedings of the 6th LREC.* Marrakech, Morocco.

Bento Carlos Dias-da-Silva, Mirna Fernanda de Oliveira, Hélio Roberto de Moraes. 2002. Groundwork for the development of the Brazilian Portuguese Wordnet. In: *Proceedings of the 3rd International Conference Portugal for Natural Language Processing* (PorTal). Faro, Portugal. Berlin: Springer-Verlag, 189-196.

Bernardo Magnini and Manuela Speranza. 2001. Integrating generic and specialized wordnets. In: *Proceedings of the Conference on Recent Advances in Natural Language Processing*. Bulgaria.

Christiane Fellbaum (ed.). 1998. *Wordnet: an electronic lexical database*. The MIT Press, Ca, MA: 423p.

Delphine Bernhard. 2006. Multilingual term extraction from domain-specific corpora using morphological structure. In: *Proceedings of the 11th European Chapter Meeting of the ACL*, Trento, Italy, 171-174.

German Rigau Claramunt. 1998. *Automatic acquisition of lexical knowledge from MRDs*. PhD Thesis. Departament de Llenguatges i Sistemes Informàtics, Barcelona.

Gladis Maria Barcellos de Almeida. 2006. A Teoria Comunicativa da Terminologia e a sua prática. *Alfa*, 50:81-97

Gladis Maria de Barcellos Almeida and Margarita Correia. 2008. Terminologia e corpus: relações, métodos e recursos. In: Stella E. O. Tagnin and Oto Araújo Vale (orgs.). *Avanços da Lingüística de Corpus no Brasil*. 1 ed. Humanitas/FFLCH/USP; São Paulo, volume 1, 63-93.

Gladis Maria Barcellos de Almeida, Sandra Maria Aluisio and Leandro H. M. Oliveira. 2007. O método em Terminologia: revendo alguns procedimentos. In: Aparecida N. Isquerdo and Ieda M. Alves. (orgs.). *Ciências do léxico: lexicologia, lexicografia, terminologia*. 1 ed. Editora da UFMS/Humanitas: Campo Grande/São Paulo, volume 3, 409-420.

John Durkin. 1994. *Expert Systems: Design and Development.* Prentice Hall International, London, 800p.

John Sinclair, J. 2005. Corpus and text: basic principles. In: Martin Wynne (ed.). *Developing linguistic corpora: a guide to good practice.* Oxbow Books: Oxford, 1-16. Available at http://ahds.ac.uk/linguistic-corpora/

Lucelene Lopes, Paulo Fernandes, Renata Vieira and Gustavo Fedrizzi. 2009. ExATOlp - an automatic tool for term extraction from Portuguese language corpora. In: *Proceedings of the LTC'09*, Poznam, Poland.

Luisa Bentivogli, Andrea Bocco and Emanuele Pianta. 2004. ArchiWordnet: integrating Wordnet with domain-specific knowledge. In: *Proceedings of the 2nd International Global Wordnet Conference*. Masaryk University, Brno, 39-47.

Luiz Carlos Ribeiro Jr. 2008. *OntoLP: construção semi-automática de ontologias a partir de textos da língua portuguesa*. MSc Thesis, UNISINOS, 131p.

Maria Fernanda Bacelar do Nascimento. 2003. O papel dos corpora especializados na criação de bases terminológicas. In: I. Castro and I. Duarte (orgs.). *Razões e emoções, miscelânea de estudos em homenagem a Maria Helena Mateus*. Imprensa Nacional-Casa da Moeda: Lisboa, volume II, 167-179.

Maria Tereza Cabré. 1999. *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Institut Universitari de Linguística Aplicada: Barcelona.

Maria Tereza Cabré, Anne Condamines and Fidelia Ibekwe-SanJuan. 2005. Application-driven terminology engineering. *Terminology*, 11(2):1-19.

Maria Tereza Cabré, Rosa Estopà and Jordi Vivaldi Palatresi. 2001. Automatic term detection: a review of current systems, In: Didier Bourigault et al. (eds.). *Recent Advances in Computational Terminology*. John Benjamins Publishing Co: Amsterdam & Philadelphia, 53-87.

Maria Teresa Pazienza, Marco Pennacchiotti and Fabio Massimo Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. *Studies in Fuzziness and Soft Computing*, 185:255-280.

Maria Teresa Sagri, Daniela Tiscornia and Francesca Bertagna. 2004. Jur-Wordnet. In: *Proceedings of the 2nd International Global Wordnet Conference*. Masaryk University, Brno, 305-310.

Marti A. Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings 14th of the International Conference on Computational Linguistics*. Nantes, 539-545.

Paul Buitelaar and Bogdan Sacaleanu. 2002. Extending synsets with medical terms. In: *Proceedings of the 1st International Global Wordnet Conference*. Mysore, India, 2002.

Piek Vossen (ed.). 2002. EuroWordnet general document (Version 3–Final). Available at: http://www.vossen.info/docs/2002/EWNGeneral.pdf.

Satanjeev Banerjee and Ted Pedersen. 2003. The Design, Implementation, and Use of the Ngram Statistics Package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City.