An Introduction to the TermiNet Project

[1,2] Ariani Di Felippo
[1] Grupo de Estudos e Pesquisas em Terminologia (GETerm)
Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
CP 676 – 13565-905, São Carlos, SP, Brazil
[2] Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
CP 668 – 13.560-970, São Carlos, SP, Brazil

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/ hyponymy, holonymy/ meronymy, entailment, and cause) 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. Many recent projects with the objective of (i) integrating generic and specialized wordnets (e.g., Magnin, Speranza, 2001; Roventini, Marinelli, 2004; Bentivogli *et al*., 2004), (ii) enriching generic wordnets with terminological units (e.g., Buitelaar, Sacaleanu, 2002) or (iii) constructing terminological wordnets (e.g., Sagri et al., 2004) have shown that lexical resources with domain-specific coverage are crucial for the development of concrete NLP applications. In other words, a concrete NLP application must be able to comprehend both expert and non-expert vocabulary. The reason behind these projects is that general semantic lexicons like WN.Pr do not cover many terms and concepts specific to certain domains therefore these resources need to be tuned to a specific domain. In all the aforementioned projects, however, there is no clear and generic methodology for building terminological wordnets (or terminets), despite the existence of a reasonable number of them. Consequently, the **TermiNet** (*"Terminological WordNet"*) project has started in 2009 (September) with the financial support of *State of São Paulo Research Foundation* (FAPESP/ Proc. 09/06262-1). At present, the project is been developed in the laboratory of *Research Group of Terminology*[1] (GETerm) in *Federal University of São Carlos* (UFSCar) with the collaboration of the *Interinstitutional Center for Research and Development in Computational Linguistics*[2] (NILC/University of São Paulo) researchers. Specifically, the TermiNet project has two goals. The first one is to instantiate the generic NLP methodology, proposed by Dias-da-Silva (2006), for developing terminets. Such methodology distinguishes itself by conciliating the linguistic and computational facets of the NLP researches. The second one is to apply the instantiated methodology to build a terminet in Brazilian Portuguese (BP), since BP is a resource-poor language in NLP for which domain-specific

[1] http://www.geterm.ufscar.br/
[2] http://www.nilc.icmc.usp.br

databases in wordnet format have not been built so far. Assuming a compromise between Human Language Technology and Linguistics, and based on the Artificial Intelligence notion of Knowledge Representation, the methodological approach proposed by Dias-da-Silva (2006) claims that the linguistic-related information to be computationally modeled, like a rare metal, must be "mined", "molded", and "assembled" into a computer-tractable system. Accordingly, the processes of designing and implementing a terminet lexical database have to be developed in the following complementary domains: (a) *the linguistic-related domain*, where the lexical resources and the lexical-conceptual knowledge are mined; the lexical resources can be dictionaries, *thesauri,* taxonomies, text *corpora*, etc., from which (i) the terminological units (or terms), (ii) the lexical relations, (iii) the conceptual relations, (iv) the glosses, and (v) the co-text sentences are extracted; (b) *the representational domain*, where 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 "synset" – a set of words built on the basis of the notion of synonymy in context, i.e. word interchangeability in some context –, and "lexical matrix" – associations of sets of word forms and the concepts they lexicalize, and (c) *the computational domain*, where the computer-tractable representations are assembled by means of utilities (i.e., a computational tool to create and edit lexical knowledge). The instantiated methodology will be validated by building DE.WordNet (DE.WN), a specialized wordnet for the Distance Education (or Distance Learning) domain in BP. The Distance Education is a rapidly growing and evolving field for which no computational semantic lexicon in BP is available. DE.WN can be integrated into the wordnet lexical database for Brazilian Portuguese, the WordNet.Br (Dias-da-Silva *et al*., 2008), enriching it with domain specific knowledge. Besides the benefits to NLP domain, terminets may also contribute to the development of terminological/terminographic products since the organization of the lexical-conceptual knowledge is an essential step in building such products. The results of TermiNet will be freely accessible in the web with the requirement of a license. The particular relevance of this project is that it will provide a generic methodology that could be applied to the construction of terminets for other domains and languages and a computational lexical resource that shall have many applications. At GLAT 2010, we intend to present more details about the project.

References

Bentivogli, I, L.; Bocco, A.; Pianta, E. (2004) ArchiWordnet: integrating Wordnet with domain-specific knowledge. In the *Proceedings of the 2nd International Global WordNet Conference*, pp. 39-47. Brno, Czech Republic. January 20-23.

Buitelaar, P.; Sacaleanu, B. (2002) Extending synsets with medical terms. In the *Proceedings of the 1st International Conference on Global WordNet.* Mysore, India. January 21-25.

Dias-da-Silva, B. C. (2006) O estudo linguístico-computacional da linguagem. *Letras de Hoje*, Porto Alegre, v. 41, n. 2, p. 103-138.

___.; Di Felippo, A.; Nunes, M.G.V. (2008) The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In the *Proceedings of the 6th International Conference on Language Resources and Evaluation* – LREC, pp. 1535-1541. Marrakech, Morocco.

Fellbaum, C., (Ed.) (1998) *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, 423 p.

Hanks, P. Lexicography. In: Mitkov, R. (Ed.). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press, 2004, p. 48-69.

Magnini, B.; Speranza, M. (2001) Integrating generic and specialized wordnets. In the *Proceedings of the 2nd Conference on Recent Advances in Natural Language Processing* – RANLP, pp. 149-153. Tzigov Chark, Bulgaria.

Roventini, A.; Marinelli, R. (2004) Extending the Italian Wordnet with the specialized language of the maritime domain. In the *Proceedings of the 2nd International Global WordNet Conference*, pp. 193-198. Brno, Czech Republic. January 20-23.

Sagri *et al*. (2004) Jur-Wordnet. In the *Proceedings of the 2nd International Global WordNet Conference*, pp. 305-310. Brno, Czech Republic. January 20-23.