REBECA: Turning WordNet Databases into "Ontolexicons"

[1,2]Bento Carlos Dias-da-Silva, [2,3]Ariani Di Felippo,
[1]Centro de Estudos Linguísticos e Computacionais da Linguagem (CELiC)
Departamento de Letras Modernas – Faculdade de Ciências e Letras – Universidade Estadual Paulista (UNESP)
CP 174 – 14.800-901, Araraquara, SP, Brasil
[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, Brasil
[3]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
CP 676 – 13565-905, São Carlos, SP, Brasil
bento@fclar.unesp.br; ariani@ufscar.br

Abstract

In this paper we outline the design and present a sample of the REBECA bilingual lexical-conceptual database constructed by linking two monolingual lexical resources in which a set of lexicalized concepts of the North-American English database, the Princeton WordNet (WN.Pr) synsets, is aligned with its corresponding set of lexicalized concepts of the Brazilian Portuguese database, the Brazilian Portuguese WordNet synsets under construction, by means of the MultiNet-based interlingual schema, the concepts of which are the ones represented by the Princeton WordNet synsets. Implemented in the Protégé-OWL editor, the alignment of the two databases illustrates how wordnets can be turned into ontolexicons. At the current stage of development, the "wheeled-vehicle" conceptual domain was modeled to develop and to test REBECA's design and contents, respectively. The collection of 205 ontological concepts worked out, i.e. REBECA´s alignment indexes, is exemplified in the "wheeled- vehicle" conceptual domain, e.g. [CAR], [RAILCAR], etc., and it was selected in the WN.Pr database, version 2.0. Future work includes the population of the database with more lexical data and other conceptual domains so that the intricacies of adding more concepts and devising the spreading or pruning the relationships between them can be properly evaluated.

1. Introduction

In knowledge-based Natural Language Processing (NLP) systems, the lexical knowledge database is responsible for providing the language lexical forms with their morphosyntactic and conceptual-semantic properties to the NLP processing modules (Hanks, 2004).

There is no doubt about the increasing need of robust and accurate general language lexical/semantic resources for developing NLP applications (Calzolari, 2004). These resources include lexicons, lexical databases, lexical knowledge bases, and ontologies. In particular, applications such as knowledge-based machine translation and multilingual information retrieval systems require bilingual and/or multilingual robust lexical resources. Outstanding multilingual initiatives are SIMPLE (Lenci et al. 2000) and EuroWordNet (EWN) (Vossen, 1998). The former is a framework for developing general multilingual lexicons anchored on a general language ontology and the latter is a WordNet-based multilingual database that connects different monolingual wordnets through equivalence relations for each language synset (i.e., synonym set) to the closest concept from the so-called Inter-Lingual-Index (ILI).

Brazilian Portuguese NLP researchers are in need of robust machine tractable bilingual and multilingual lexical resources. In this scenario, we can highlight the Brazilian Portuguese WordNet (WordNet.Br or WN.Br) initiative, which is being linked to Princeton WordNet (WN.Pr), version 2.0, basically along the lines as the EWN project (Dias-da-Silva et al. 2008); another initiative, relying on a complementary method of building bilingual databases, is the design and construction of the REBECA lexical database (Di-Felippo & Dias-da-Silva, 2008). The design of this "ontolexical" resource is the result of a PhD study on lexicalization mismatches between Brazilian Portuguese (BP) and North-American English (AmE) (Di-Felippo, 2008).

Accordingly, this paper aims to outline REBECA's architecture and to show its relevance to aid linguists in the tasks of analyzing cross-language lexicalization patterns and compiling lexical information paired with its conceptual counterpart. In Section 2 we sketch out REBECA's architecture. In Section 3 we outline its methodological underpinning and its constructional steps. In Section 4 we illustrate how REBECA registers AmE/BP lexicalization mismatches. In Section 5 concluding remarks close the paper.

2. The REBECA lexical database

As a sort of "dual" lexical database, REBECA aligns a set of lexicalized concepts (i.e. concepts that are linguistically expressed by means of synsets) of the WN.Pr database with its corresponding set in the WN.Br database by means of a specific ontology, i.e. its mapping interlingua.

At the current stage of development, the "wheeled-vehicle" conceptual domain was modeled to develop and to test REBECA's design and contents, respectively. The domain gathers those concepts that represent "concrete entities" that are linguistically expressed by nominals, e.g. the concept [CAR] is expressed in AmE by the synset {car, auto, automobile, machine, motorcar} and in BP by the synset {auto, automóvel, carro}.

3. The methodology

Assuming a compromise between NLP and Linguistics, and based on the Artificial Intelligence notion of Knowledge Representation Systems (Durkin, 1994), REBECA is developed within the three-domain approach methodology (Dias-da-Silva et al., 2008) that claims that the linguistic-related information to be computationally modeled must be "mined", "molded", and "assembled" into a computer-tractable system. Accordingly, the processes of designing and implementing the REBECA lexical database fall within the following complementary domains:

- *The Linguistic-related Domain*, where the lexical resources (dictionaries and text corpora), the ontology, and the lexicalized concepts (i.e., lexical units in the synsets) are mined (Handke, 1995);
- *The Computational Linguistic-related Domain*, where the ontology concepts and their linguistic expressions in AmE and BP are molded into a computer-tractable representation language; specifically, the formal representation of the ontological concepts in REBECA adapts the Multilayered Extended Semantic Networks (MultiNet), a specific knowledge representation formalism which provides the semantic representatives for the description of the semantics of natural language expressions (Helbig, 2006);
- *The Computational-related Domain*, where the computer-tractable representations are assembled by means the Protégé-OWL editor (Horridge, 2004) and its TGVizTab plug-in (Alani, 2003).

3.1. The Linguistic-related Domain

a) The reference corpus

Given the unavailability of reusable machine-readable BP dictionaries and other resources, the REBECA developers manually reused, merged, and tuned lexical-conceptual information registered in: (i) two bilingual AmE-BP dictionaries (Houaiss & Cardim, 1982; Weiszflog, 2000); (ii) two synonyms (and antonyms) dictionaries in PB (Barbosa, 2000; Fernandes, 1997); (iii) two monolingual dictionaries of BP (Houaiss & Villar, 2001; Ferreira, 2004); (iv) the WN.Br lexical database, and (v) BP texts available in the PLN-BR FULL *Corpus* and web.

The PLN-BR FULL is a 29 million words *corpus* of contemporary BP; specifically, it is a journalistic *corpus* that is comprised of articles taken from the Brazilian newspaper Folha de São Paulo. The PLN-BR FULL is available for online search at the Philologic webpage[1]. Texts in BP available on the web were "googled".

b) The ontology concepts

The collection of the ontological concepts (i.e. the alignment indexes) is exemplified in the "wheeled-vehicle" domain (e.g., [CAR], [RAILCAR], etc.)[2] and was selected in the WN.Pr 2.0 database. Manually, we compiled 205 concepts that were more specific that the concept encoded in the synset {wheeled-vehicle}, i.e. its hyponyms. Each ontological concept (a sort of ILI) was identified by means of a lexical unit and by a BP gloss that specified its underlying concept.

a) The AmE monolingual database lexical units

The 205 lexicalized concepts in AmE database were extracted from the WN.Pr 2.0 database. It should be noted that for each word-form in a synset, one co-text sentence (i.e. a sentence that provides the context of minimal use) was manually extracted from WN.Pr 2.0 or web. The web co-text sentences were selected with the WebCorp[3] search engine.

b) The BP monolingual database lexical units

REBECA's BP database stores 84 lexicalized concepts of the "wheeled-vehicle" domain.

Specifically, the manual process of identifying the BP expressions started out with a specific WN.Pr 2.0 synset. Then, it was identified the BP equivalences for all word-forms of the WN.Pr synset with the help of the bilingual AmE-BP dictionaries, which provided equivalent BP expressions, both lexical units (simple, compound or complex lexical units) or "recurrent free phrases"[4] (RFPs).

When the expressions were lexical units, we followed the steps described below:

(i) Bilingual dictionary look up to initiate the BP synset *x* cyclic construction;

(ii) WN.Br synset database look up to feed the synset *x* and to check its concept encoding soundness. This step outputs the synset *x'*, the first reformulation of the synset *x*;

(iii) Synonym dictionary look up to feed the synset *x'*. This step outputs the synset *x''*, the second reformulation of the synset *x*;

(iv) Monolingual dictionary look up to feed the synset *x''*. This step outputs the synset *x'''*, the third reformulation of the synset *x*;

(v) The PLN-BR FULL corpus and the web look up to check the use of the lexical units of the synset *x'''*. This step outputs the synset x'''', the fourth reformulation of the synset *x*, i.e. its final version.

When the expression looked up in the bilingual dictionaries was an RFP, we followed two specific steps: (i) the construction of an initial with the RFPs entered in the AmE-BP dictionaries, and (ii) the corpus look up to find occurrences of equivalent RFPs.

[1] http://philologic.uchicago.edu/

[2] The encoding of concepts from the "wheeled-vehicle" domain is just the starting point of REBECA database construction. Events and states are planned are to be dealt with in the future.

[3] WebCorp is a fully tailored linguistic search engine to cache and process large sections of the web (http://www.webcorp.org.uk).

[4] See Bentivogli & Pianta (2004).

The identification process finished, 121 lexical gaps in the "wheeled-vehicle" domain were spotted in BP lexicon, 40 of which are filled in by RFPs (Bentivogli & Pianta, 2004). Thus, the RFPs work as an alternative way to express non-lexicalized concepts. For example: as the concept [GOLF_CART] is not lexicalized in BP it is expressed by the RFP *carrinho de golfe*.

3.2. The Representational Domain

a) The ontology representation

The choice of MultiNet was motivated for (i) its means of representation are powerful enough to express the concepts underlying lexical expressions, and (ii) every concept has a single representation through which all information associated with it becomes available. Furthermore, MultiNet has already been successfully used to construct computational lexicons. Specifically, the MutiNet is the semantic representation formalism of the HaGenLex (HAgen GErmaN LEXicon), a domain independent computational lexicon for German (Hartrunpft et al, 2003). In fact, the semantic representation in REBECA database differs from HaGenLex because its focuses on the description of the superordinate relations between the concepts (lexical meanings), without any description of syntactic-semantic aspects of the lexical meaning (e.g. argument structure). An overview of MultiNet's representational means is given in Figure 1[5].

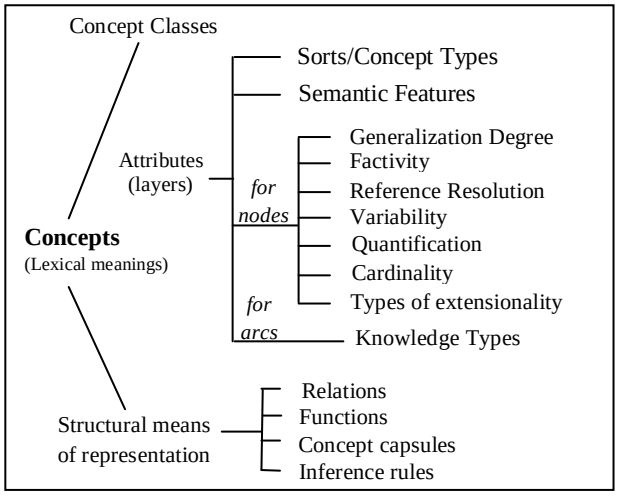

Figure 1: MultiNet semantic representational means.

According to MultiNet, the ontology concepts (the network nodes) of RECEBA are organized by means of superordinate relations (the network arcs), which are labeled as SUB (i.e. subsumption), e.g. [WHEELED VEHICLE] SUB [BICYCLE]). Moreover, the ontology concepts are linked to one another by means of PARS (i.e. part-whole relation), e.g. [BICYCLE] PARS [WHEEL]), and PURP (i.e. purpose) relations, e.g., [BICYCLE] PURP [LOCOMOTE]). It is important to mention that the MultiNet formalism defines a fixed set of about 140 semantic relations and functions, but only three relations were used in the development of REBECA, since SUB, PARS, and PURP are the most relevant relations to organize nominal concepts.

However, a concept related by PARS (e.g. [WHEEL]) or PURP (e.g. [LOCOMOTE]) is not an ontology index; it is a property of the ontology concept. The encapsulation of concepts ensures that the knowledge established by a type of relation be properly inherited by the more specific concepts: e.g. the [WHEELED VEHICLE] is linked to [AIR BAG] by means of PARS, then its hyponyms inherit the PARS relation. This is because PARS is considered a prototypical knowledge, which is inherited as default knowledge in the conceptual hierarchy.

The classificatory means are based on sorts (or ontology classes), semantic features, and multidimensional attributes, which are responsible for encoding intensional and extensional aspects of meanings.

According to the taxonomic means, the ontological sort of concepts like [CAR] is [SORT=movable-artifact-discrete] and the corresponding semantic features are [artif+], [instru+], and [movable+]. Thus, every concept of the Interlingua is associated with them.

As the ontology indexes are generic concepts (e.g. [BICYCLE]), they are specified by the following attribute-value pairs: [GENER=ge], [REFER=refer], [VARIA=con], [FACT=real], and [ETYPE=0]. The value "ge" of the multidimensional attribute GENER indicates a generic concept. For generic concepts the value of the attribute REFER remains underspecified, which is represented as [REFER=refer]. The value "con" of the attribute VARIA indicates that a concept like [BICYCLE] is a fixed element in the extensional level. The value "real" of the attribute FACT indicates a real object in the extensional level. The value "0" indicates that the extensional of a generic concept "x" is a prototypical element of the set <all x>. The attribute of the arcs labeled by SUB is "K" (i.e. categorical knowledge) and by PARS or PURP is "D" (i.e. default knowledge).

The essential distinction between these two knowledge types is related to the inferential processes. The K knowledge of the scope of meaning of a generic concept is strictly inherited (i.e. without exceptions) by all sub-concepts and subordinated specializations (the carrier of this inheritance are the subordination relations SUB). The prototypical knowledge of a generic concept´s scope of meaning, on the other hand, has to be taken as D knowledge, which is inherited from top to bottom in the hierarchy of concepts, similar to the way K concepts. In this context, the term "default" denotes a basic property that is valid as long as there is no other information available In contrast to K knowledge, D knowledge can be revised or overwritten in exceptional cases (Helbig, 2006).

[5] See more information about MultiNet formalism at http://pi7.fernuni-hagen.de/research/multinet/multinet_en.html.

b) The linguistic representation
Following the wordnet format (Fellbaum, 1998), the lexicalized concepts in both monolingual databases are encoded as synsets. The RFPs are encoded as phrasets (i.e. a set of RFPs) (Bentivogli & Pianta, 2004), e.g. {carrinho de golfe}.

3.3. The Computational-related Domain

The REBECA database is being constructed with the help of the Protégé-OWL editor. The "wheeled-vehicle" domain lexical-conceptual data were typed in as follows:

(i) The ontology concepts were entered as "classes" of Protégé-OWL;

(ii) Other concepts, which are linked to the ontology concepts by PARS and PURP relations, the sort, the semantic features, and the multidimensional attributes are typed in as "properties" of the classes;

(iii) The synsets of each monolingual database (AmE and BP databases) and the phrasets of the BP database were typed in as "instances" or "individuals" of the classes, i.e. language-specific synsets that are instances of the same ontology index are thus equivalent across the two languages; in the cases of lexical gaps (not filled with phrasets) it was possible to traverse the ontology structure in search for more generic lexicalizations in BP that could be used as alternative expressions for the non-lexicalized concepts;

(iv) The glosses were entered as "comments" of the classes; and

(v) The co-text sentences were entered as "comments" of lexical units or SLRs.

The Figure 2 exemplifies the alignment of the two databases in the Protégé-OWL Editor, a sort of "ontolexicon" (Prévot, Borgo, and Oltramari, 2005).

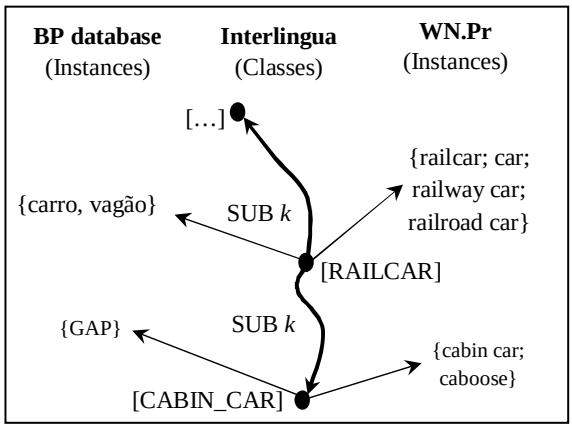

Figure 2. Lexical-conceptual alignments in REBECA.

4. Lexicalization mismatches

After identifying the lexicalized concepts in BP, it is possible to observe lexicalization mismatches between AmE and BP regarding the "wheeled-vehicle" conceptual domain. To illustrate these mismatches, consider Figure 3, where the hierarchical structure extracted from WN.Pr 2.0 reflects the combination of lexicalized synsets and non-lexicalized (in capital letters) concepts in AmE. Specifically, it can be seen that in AmE the concept [WAGON] has many hyponyms (e.g., [COVERED_WAGON], [CART], etc). This concept is not lexicalized in BP, resulting in a flatter lexical-conceptual system (Figure 4). The concepts [COVERED_WAGON] and [CART], which are lexicalized in BP by {carroção} and {carroça}, respectively, are on the same level as {bicicleta; bike; magrela} in the BP hierarchy. In other words, the synsets {carroção} and {carroça} are direct hyponyms of the concept [WHEELED_WAGON].

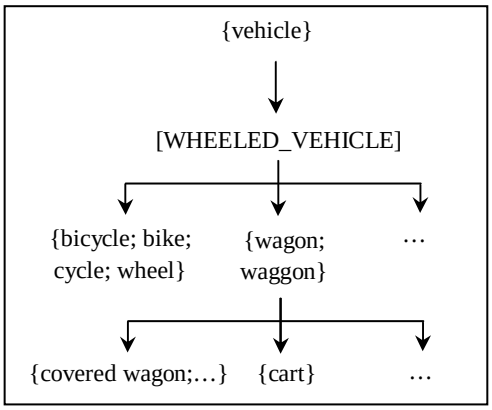

Figure 3: Lexical-conceptual organization in AmE.

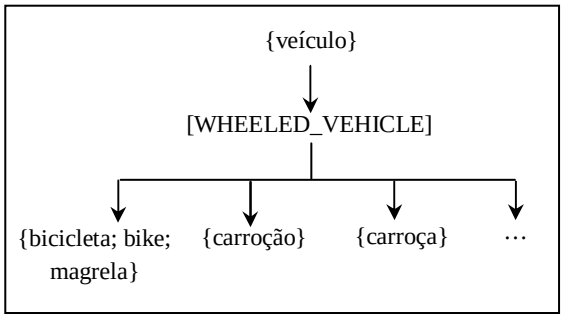

Figure 4: Lexical-conceptual organization in BP.

5. Final Remarks

This paper presented REBECA's architecture and showed its relevance as a specific resource for developing linguistically motivated ontolexicons, as it aids the linguist in the hard task of analyzing and linking lexical and conceptual information into a database by means of a rich MultiNet-based interlingual schema. The database, though, needs to be populated with more lexical data and more domains so that the intricacies of adding more concepts and devising the spreading or pruning the relationships between can be properly evaluated. That is the work on the way!

6. Acknowledgements

The authors are grateful to CNPq (Brazilian National Council for Scientific and Technological Development) for supporting the PhD research during which the REBECA lexical database was designed, to the LREC´2010 referees, who helped make this paper better, and to UNESP-PROPG and CAPES for making the participation in this event possible.

7. References

Alani, H. (2003) TGVizTab: an ontology visualisation extension for Protégé. In: *Proceedings of the Workshop on Visualization Information in Knowledge Engineering (VIKE´03)*. Sanibel Island, Florida, USA.

Barbosa, O. (2000). *Grande dicionário de sinônimos e antônimos*. Rio de Janeiro: Ediouro.

Bentivogli, L.; Pianta, E. (2004). Extending wordnet with syntagmatic information. In *Proceedings of the 2nd Global Wordnet Conference* (GWC´04), Czech Republic, pp. 47–53.

Calzolari, N. (2004). Computational lexicons and corpora: complementary components in human language technology. In P. van Sterkenburg (Ed.), *Linguistics Today: facing greater challenge*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 89-107.

Dias-da-Silva, B.C.; Di Felippo, A.; Nunes, M.G.V. (2008). The automatic mapping of Princeton WordNet lexical-conceptual relations onto the Brazilian Portuguese WordNet database. In *Proceedings of the 6th Language Resources and Evaluation Conference* (LREC´08). Marrakech, Morocco, pp. 335-342.

Di-Felippo, A; Dias-da-Silva, B. C. (2008). REBECA: uma base de dados léxico-conceituais bilíngüe inglês-português. In *Proceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence* (WTDIA/SBIA´08), Salvador-BA, Brazil.

Durkin, J. (1994). *Expert Systems: design and development*. London: Prentice Hall International.

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

Fernandes, F. (1997). *Dicionário de sinônimos e antônimos da língua portuguesa*. São Paulo: Globo.

Ferreira, A. B. H. (2004). Novo dicionário eletrônico Aurélio da língua portuguesa. Curitiba: Ed. Positivo. 1 CD-ROM

Hanks, P. (2004) Lexicography. In R. Mitkov (Ed.). *The Oxford handbook of Computational Linguistics*. Oxford: Oxford University Press, pp. 48-69.

Handke, J. (1995). *The Structure of the Lexicon: human versus machine*. Berlin: Mouton de Gruyter.

Hartrumpf, S.; Helbig, H.; Osswald, R. (2003). The semantically based computer lexicon HaGenLex - Structure and technological environment. *Traitement Automatique des Langues*, 44(2):81-105.

Helbig, H. (2006) *Knowledge representation and semantics for natural language*. Berlin/Heidelberg: Springer-Verlag.

Horridge, M. et al. (2004) *a practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools*. The University Of Manchester. Available at <http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf>.

Houaiss, A.; Villar, M. de S. (2001). *Dicionário eletrônico Houaiss da língua portuguesa*. (versão 1.0). Rio de Janeiro: Editora Objetiva. CD-ROM.

Houaiss, A.; Cardim, I. (Orgs.). (1982). *Dicionário eletrônico Webster's inglês-português/ português-inglês*. Rio de Janeiro, Ed. Record, 1982. 1 CD-ROM

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). SIMPLE: a general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4), pp. 249–263.

Prévot, L.; Borgo, S.; Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. In: *Proceedings of OntoLex 2005*. Jeju Island, Korea.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, 32 (2-3), pp. 73-89.

Weiszflog, W. (2000). *Michaelis: moderno dicionário inglês (inglês-português/ português-inglês)*. Editora Melhoramentos. Disponível em <http://michaelis.uol.com.br/moderno/ingles/index.php>.