IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 137-150
RECEIVED 10/11/09 ACCEPTED 16/01/10 FINAL 10/03/10

Brazilian Portuguese WordNet: A Computational Linguistic Exercise of Encoding Bilingual Relational Lexicons

BENTO CARLOS DIAS-DA-SILVA

Universidade Estadual Paulista (UNESP), Brazil

ABSTRACT

This paper describes the methodology of encoding the Brazilian Portuguese WordNet (WN.Br) synsets and the automatic mapping of WN.Br´s conceptual relations of hyponymy, co-hyponymy, meronymy, cause, and entailment relations from Princeton WordNet (WN.Pr). After contextualizing the project and outlining the current lexical database structure and its statistics, it is described the WN.Br editing tool to encode the synsets, its glosses and the equivalence EQ_RELATIONS between WN.Br and WN.Pr synsets, and to select sample sentences from corpora. The conclusion samples the automatic generation of WN.Br´s hyponymy and co-hyponymy conceptual relations from WN.Pr and outlines the ongoing work.

1 INTRODUCTION

Natural language processing (NLP) initiatives to design, build, and compile precise, rich, and robust lexicons for NLP applications are extremely time-consuming and prone to flaws tasks [1] [2], [3] due to the fact that lexicon developers are expected to specify and code huge amounts of specialized and interrelated information as phonetic/graphemic, morphological, syntactic, semantic, and even illocutionary bits of information into computational lexicons [4].

Princeton WordNet (WN.Pr), for example, is a successful sort of a computational lexicon that has set the pattern for compiling bulky relational lexicons systematically. An on-line relational lexical semantic database, WN.Pr combines the designs of a dictionary and of a thesaurus. Similar to a standard dictionary, it covers nouns, verbs, adjectives, and adverbs. After 18 years of research, its 1998 database version (v. 1.6) contained about 94,000 nouns, 10,000 verbs, 20,000 adjectives, and 1,500 adverbs [5].[1] Similar to a thesaurus, words are grouped in terms of lexicalized concepts, which are, in turn, represented in terms of synonym sets (*synsets*), i.e. sets of words of the same syntactic category that share the same concept. Its web structure makes it possible for the user to find a word meaning in terms of both the other words in the same synset and the relations to other words in other synsets as well. Essentially, WN.Pr is a particular semantic network and its sought-after NLP applications have been discussed by the research community [6], [7].

Mirroring WN.Pr´s construction methodology, wordnets of other languages have been under development. EuroWordNet (EWN) [8] is the outstanding multilingual initiative. It is a multilingual wordnet that results from the connection of individual monolingual wordnets by means of encoding the equivalence EQ-RELATIONS (see section 3) between each synset of each individual wordnet and the closest concept represented by the so-called Inter-Lingual-Index (ILI)[2], which enables cross-lingual comparison of words, synsets, concept lexicalizations, and meaning relations from different wordnets [9].

Mirroring both WN.Pr's and EWN's initiatives, and extending the Brazilian Portuguese Thesaurus [10], [11], the Brazilian Portuguese WordNet (WN.Br) project was launched in 2003 and the WN.Br database has been under construction since then. In particular, this paper focuses on the coding of the following bits of information in the database: (a) the co-text sentence for each word-form in a synset; (b) the synset gloss; and (c) the relevant language-independent hierarchical conceptual-semantic relations of hypernymy[3], hyponymy[4], meronymy

[1] The current version (v. 3.0) contains 101,863 nouns, 11,529 verbs, 21,479 adjectives, and 4,481 adverbs. See more details at `http://wordnet.princeton.edu`.

[2] The ILI is an unordered list made up of each synset of the WN.Pr with its gloss (an informal lexicographic definition of the concept evoked by the synset).

[3] The term Y is a hypernym of the term X if the entity denoted by X is a kind of entity denoted by Y.

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

(part-whole relation), entailment[5] and cause[6] between synsets. Accordingly, section 2 describes the current WN.Br database and its editing tool, an editing GUI (Graphical User Interface), designed to aid the linguist in carrying out the tasks of constructing synsets, selecting co-text sentences from corpora, writing synset glosses, specifying the EQ-RELATIONS, and generating the alignments between the two databases. Section 3, after addressing the issues of cross-linguistic alignment of wordnets by means of the ILI, describes the conceptual-semantic alignment strategy adopted to link WN.Br synsets to WN.Pr synsets by means of the editing tool. Section 4 concludes the paper by exemplifying the automatic mapping of the WN.Pr verb hyponymy and co-hyponymy relations onto the WN.Br verb synsets.

2 THE WORDNET.BR LEXICAL DATABASE

Currently, the WN.Br database contains 44,000 word-forms and 18,500 synsets: 11,000 verbs (4,000 synsets), 17,000 nouns (8,000 synsets), 15,000 adjectives (6,000 synsets), and 1,000 adverbs (500 synsets) [12]. The WN.Br project development strategy assumes a compromise between NLP and Linguistics and, based on the Artificial Intelligence notion of Knowledge Representation [13], [14], applies a three-domain approach methodology to the development of the database. This methodology 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 [15]. Accordingly, the process of implementing the WN.Br database is developed in three complementary domains: (a) in *the linguistic domain*, the lexical resources (dictionaries and text corpora), the set of lexical and conceptual-semantic relations, and some sort of “natural language ontology of concepts” (e.g. the "Base Concepts" and "Top Ontology" [16]) are mined; (b) in *the computational-linguistic domain*, the overall information that was selected and organized in the preceding domain is molded into a computer-tractable representation (e.g. the "synsets", the "lexical matrix", and the wordnet "lexical database" itself [5]); (c) in *the*

[5] 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 done

[6] The action A1 denoted by the verb X is the cause of the action A2 denoted by the verb Y.

computational domain, the computer-tractable representations are assembled by means of the WN.Br editing tool.

2.1 The Linguistic Domain

The WN.Br database architecture conforms to the two key representations of the WN.Pr [5]: the *synset* and the *lexical matrix*: synsets are understood as sets of word-forms built on the basis of the notion of "synonymy in context", i.e. word-form interchangeability in some context [17][7]; the lexical matrix [18] is intended to capture the many-to-many associations between word-form and meaning, i.e. the association of a word-form and the concepts it lexicalize. The lexical matrix is built up by associating each word-form to the synsets to which it is a member. Thus, a polysemous word-form will belong to different synsets, for each synset is intended to represent a single lexicalized concept.

The WN.Br synset developers (a team of three linguists) reused, merged, and tuned synonymy and antonymy information registered in five outstanding standard dictionaries of Brazilian Portuguese (BP) manually ([19], [20], [21], [22], [23, 24])[8], for there are no Brazilian Portuguese machine readable dictionaries (MRDs) and other computer tractable resources available. The NILC Corpus[9] and BP texts available in the web complemented the corpus.

2.2 The Computational-Linguistic Domain

The WN.Br database structures in terms of two lists: the List of Headwords (LH), i.e. the list of word-forms arranged alphabetically, and the List of Synsets (LS) (see Fig.1). Each WN.Br word-form belongs to the LH and is associated to a Sense Description Vector (SDV). Each SDV is co-indexed by three pointers: the "synonymy pointer", which identifies a particular synset in the LS; the "antonymy

---

[7] Antonymy, on the other hand, is checked either against morphological properties of words or their dictionary lexicographical information.

[8] The dictionaries were chosen for their pervasive use of synonymy and antonymy to define word senses, which dictated the strategy to construct the synsets by examining the dictionaries alphabetically, instead of working out synsets by semantic fields.

[9] CETENFolha. Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo. See more details at `http://www.linguateca.pt/`.

pointer", which identifies a particular antonym synset in the LS; and the "sense pointer", which identifies a particular word-form sense number in the SDV. Given such an underlying structure, each synset is linked to its gloss via the “gloss link”, and each word-form is linked to its co-text sentence via the “co-text sentence link”.

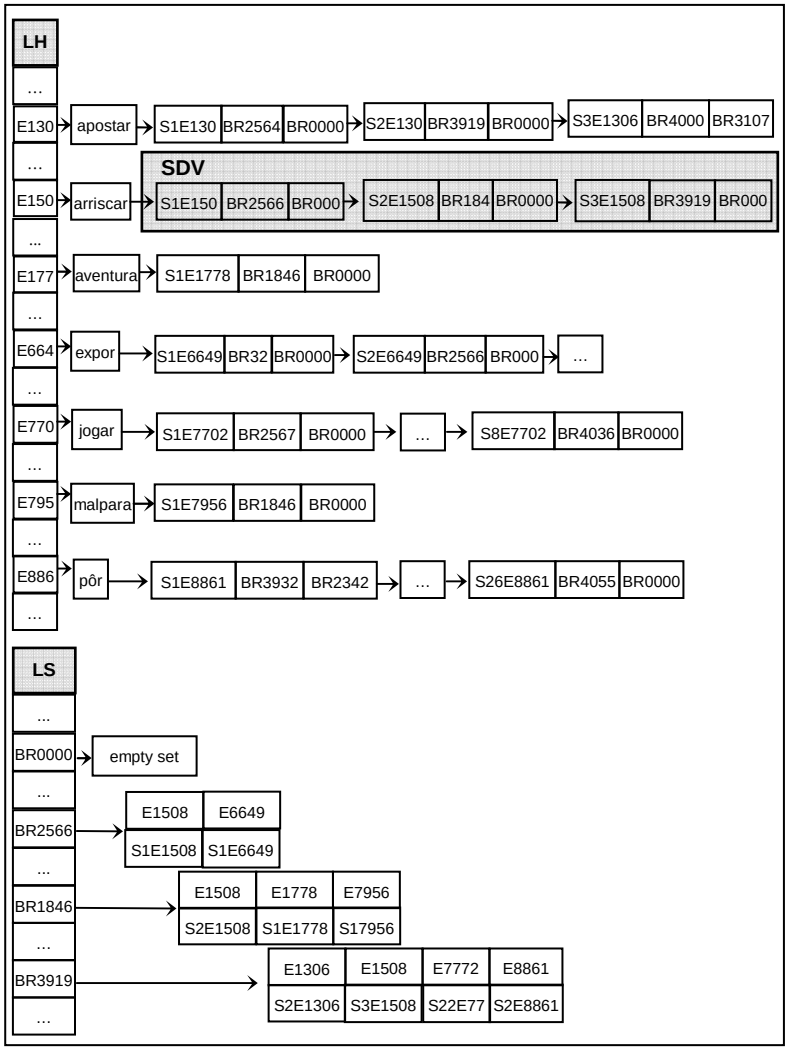

Figure 1: The WN.Br database structure.

2.3 The Computational Domain

The WN.Br editing tool is a Windows®-based GUI that allows the developers (a) to create, consult, modify, and save synsets, (b) to include co-text sentences for each word-form, (c) to write a gloss for each synset (d) to align equivalent synsets equivalence EQ-RELATIONS, (e) to code hyponymy and co-hyponymy relations in the WN.Br automatically, and (f) to generate synset lists (arranged by syntactic category, by number of elements, by the degree of homonymy and polysemy, and by co-text sentence) and WN.Br statistics. Its main functionalities include the storage and bookkeeping of the general information of the database.

The processes of using the editor can be better understood by an example. Fig. 2 shows the basic steps of constructing synsets that contain the BP verb "*lexicalizar*" ("to lexicalize"). In the first dialogue box, the developer selects the appropriate syntactic category and the expected number of synsets to be constructed (i.e. the number of senses); then, s/he clicks on the **Avançar** ("Next") button. In the second dialogue box, the **Todas as Unidades** ("All Unities") field pops up with a list of the word-forms in the WN.Br database. To construct the synset, the developer now selects the appropriate word-forms from the list and clicks on the **Avançar** button. In the third dialogue box, s/he concludes the synset construction by clicking the FIM ("End") button.

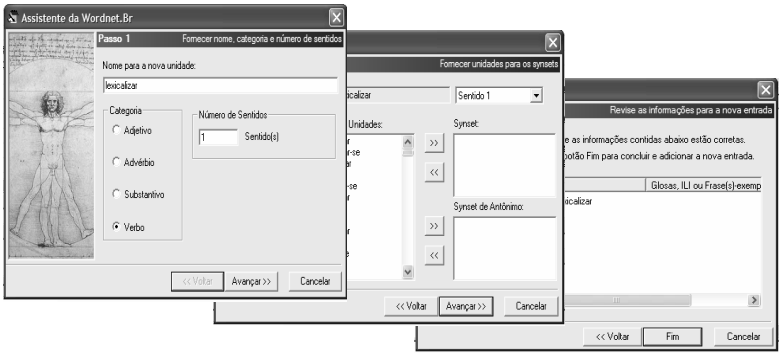

Figure 2: The synset coding wizard.

Co-text sentences, glosses, and ID numbers (see note 10) are pasted/typed in directly in the editor appropriate fields. In Fig.3, the large ellipsis highlights the **Frase(s)-exemplo** ("Sample sentences", i.e

the co-text sentences) field, and the small ellipsis, the **Glosa** ("Gloss") field and the ID number. Currently, the WN.Br database contains 19,747 co-text sentences: Table 1 shows the co-text sentence sources; Table 2 shows the number of co-text sentences per synset.

Table 1: Co-text sentence sources

Source	Nº of Co-text sentences
NILC Corpus	7,659
Aurélio [19]	732
Houaiss [25]	1,761
Michaelis [20]	858
Web	8,052
unknown	685
Total	19,747

Table 2: Co-text sentence statistics

Co-text sentences per synset	Synsets
1	18,604
2	521
3	10

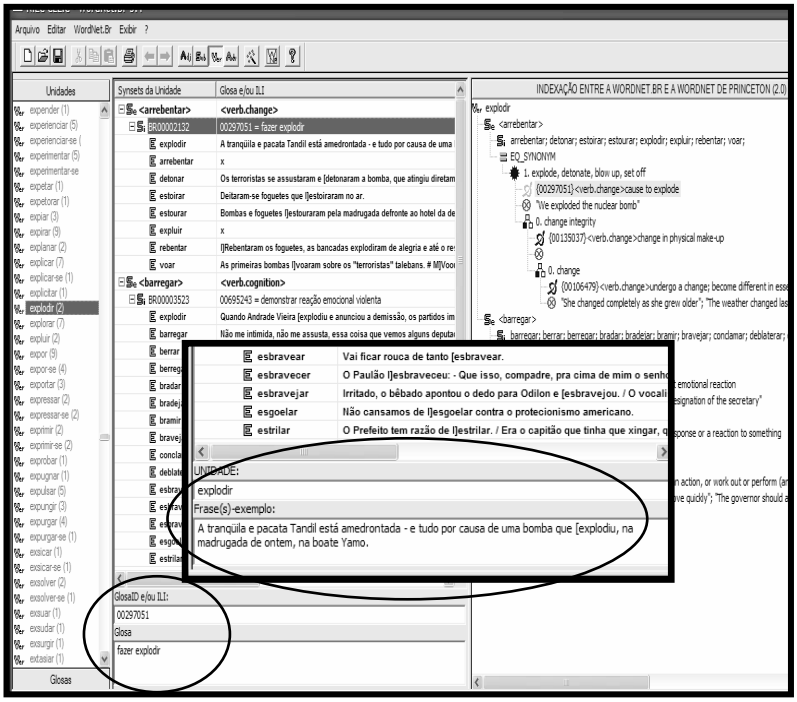

Figure 3: A screenshot with a sample of co-text sentences, glosses, ID alignment) numbers

3 CROSS-LINGUAL ALIGNMENT AND THE WN.BR CONSTRUCTION

The challenge to the WN.Br project has been to specify the equivalence EQ-RELATIONS between WN.Br and WN.Pr (v. 2.0) synsets, for such an alignment is the one that allows researchers to investigate the differences and similarities in the lexicalization processes between BP and English, to develop an English-BP lexical database which can be used in applications such as machine translation systems and cross-language information retrieval involving both languages, and to generate two types of MRDs: a monolingual BP MRD and a bilingual English-Portuguese MRD [12]. Furthermore, and most import for wordnet developers, such an alignment makes it possible the (semi-)automatic specification of the relevant conceptual-semantic relations (e.g. HYPONYMY, TROPONYMY, CO-HYPONYMY, etc.) in the wordnet under construction. In particular, in the WN.Br project, the strategy has been tested successfully to generate such hyponym and co-hyponym relations in the WN.Br verb database (see Fig 6).

The cross-lingual equivalence relations between wordnets are mined in accordance with the types identified in [8], the so-called, self defining EQ-RELATIONS (EQ-SYNONYM, EQ-NEAR-SYNONYM, EQ-HAS-HYPERONYM, and EQ-HAS-HYPONYM). Linguistic mismatches (lexical gaps, due to cultural specificities, pragmatic differences, and morphological mismatches; over/under-differentiation or of senses; and fuzzy-matching between synsets) and technical mismatches (mistakes in the choice of the appropriate EQ-RELATIONS) as have been described in [9] are also accounted for during the alignment procedure. The equivalence EQ-RELATIONS and cross-lingual mismatches are molded into a computer-tractable representation of the ILI-records[10]. The ILI-record is handy for the development, maintenance, future expansion, and reusability of a multilingual wordnet, dispenses with the development and maintenance of huge and complex semantic structures to gather all the senses encoded by each individual wordnet into a multilingual wordnet, and makes the task of adding individual wordnet to a multilingual wordnet less costly [9].

As shown in Fig.4, the structure of the WN.Br database has been extended to encode the cross-lingual equivalence EQ-RELATIONS. Besides the LH and LS lists and the SDV pointers (see 2.2), each synset structure has been augmented with an additional vector to register both the wordnet standard language - independent conceptual - semantic relations (e.g. HYPONYMY,

[10] An ILI-record is a WN.Pr (v. 2.0) synset, its gloss and its ID number [9].

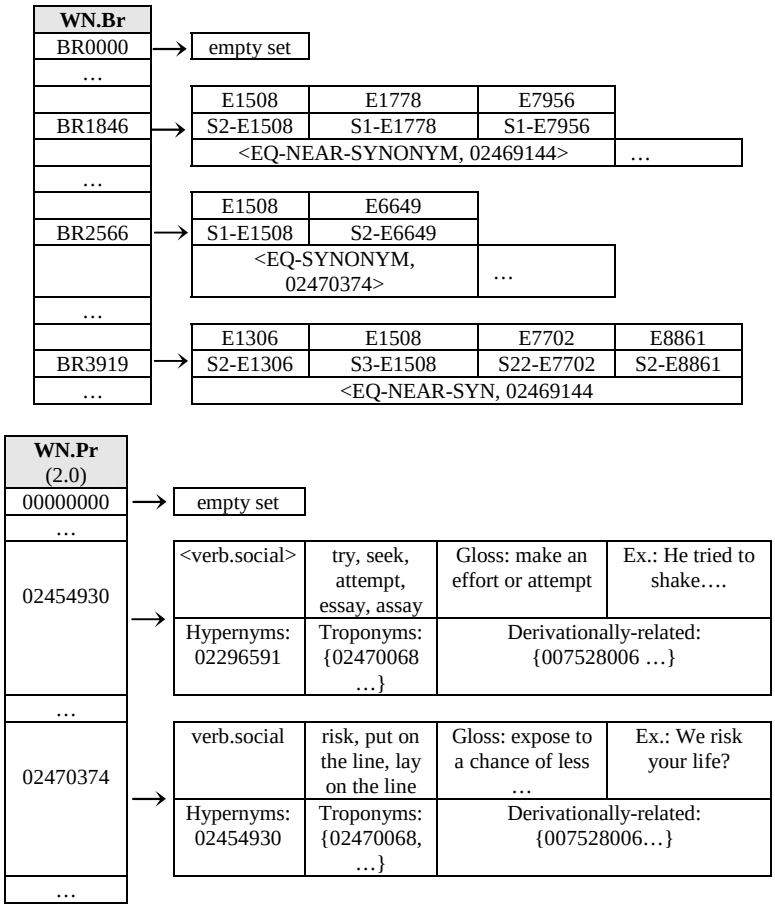

Figure 4: The synset structure augmented with conceptual-semantic EQ-RELATIONS.

TROPONYMY, CO-HYPONYMY, etc.) and the cross-lingual conceptual-semantic EQ_RELATIONS between synsets of the two wordnets. This new vector enriches the WN.Br database structure with the following cross-linguistic information: the "universal" synset semantic type (e.g. <verb.social>), the corresponding English synset (e.g. {*risk, put on the line, lay on the line*}), the English version of the universal gloss (e.g. Expose to a chance of loss or damage), the English co-text sentence (e.g. "Why risk your life?"), and EQ-RELATIONS (e.g. EQ-SYNONYM relation).

The current WN.Br editing tool has three interconnecting modules implemented as a GUI. Each module, in turn, makes it possible for the developer to carry out specific tasks during the procedure for aligning the synsets across the two wordnets: searching the BP-English dictionary, the WN.Br and WN.Pr databases, and the web.

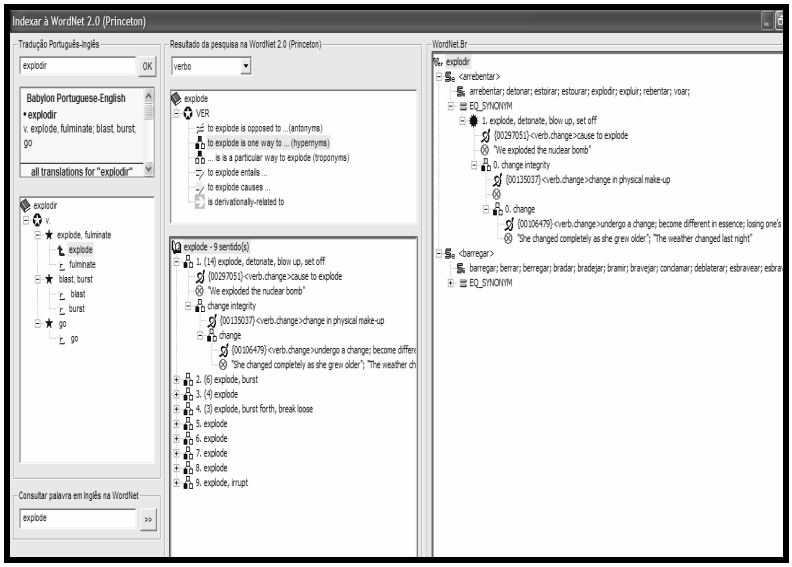

Figure 5: A screenshot of the three-column GUI of the WN.Br tool.

The WN.Br database developer starts off the alignment by right clicking on a target WN.Br word-form. As shown in Fig. 5, the editor in turn displays its three column GUI: on its left, an online bilingual BP-English dictionary and a WN.Pr database search field; in the middle, the selected WN.Pr synset information; on its right, the WN.Br synsets that contain the target word-form. The developer, in the left column, (i) checks all possible English word-forms (e.g. *explode, fulminate, blast, burst, go*) that are equivalent to the target BP word-form (e.g. *explodir*) with recourse to the dictionary and selects the appropriate one (e.g. *explode*); in the middle and right columns, (ii) analyzes the possible types of equivalence EQ-RELATIONS between the two sets of synsets: the ones in the middle column – the sets of synsets of the WN.Pr database (e.g. {*explode, detonate, blow up, set off*}, {*explode, burst*}, etc.) – and the ones in the right column – the sets of synsets of the

WN.Br database that contain the target word-form (e.g. {*arrebentar, detonar, estoirar, estourar, explodir, expluir, rebentar, voar*}, and {*barregar, berrar, berregar, bradar, bradejar, bramir, bravejar, condamar, deblaterar, esbravear, esbravejar,* }). In this particular example, the resulting EQ-SYNONYM alignment is {*explode, detonate, blow up, set off*} and {*arrebentar, detonar, estoirar, estourar, explodir, expluir, rebentar, voar*}

After the specification of alignments such as the one above, Fig. 6 sketches how the WN.Br verb database inherits both hyponym and co-hyponym relations from de WN.Pr verb database automatically. After the manual specification of the following EQ-SYNONYM alignments[11] **tentar=try**, **apostar=gamble**, and **arriscar=risk**, the WN.Br editing tool generates the following relations automatically: **apostar** and **tentar**, **arriscar** and **tentar** are hyponyms; **arriscar** and **apostar** are co-hyponyms.

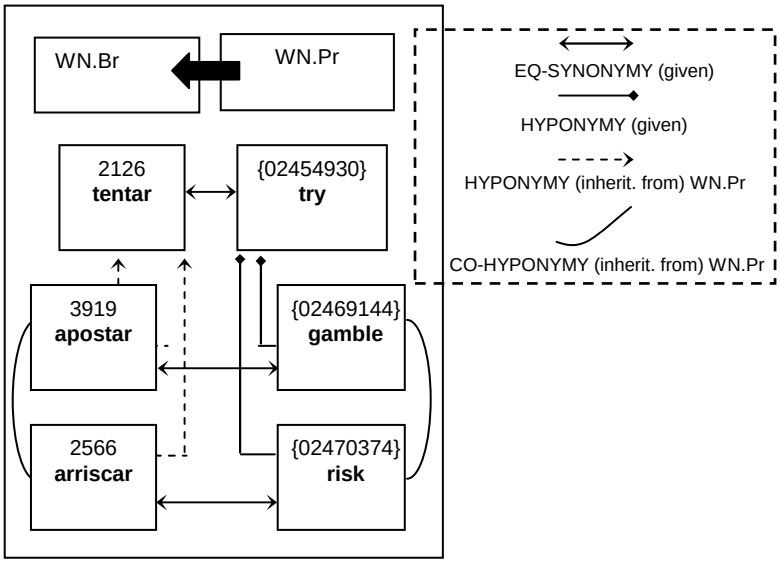

Figure 6: A sample of the automatic encoding of hyponymy and co-hyponymy relations.

[11] For short, Fig. 6 specifies the most representative **word-forms** of each synset: **tentar**: {***tentar***, *ensaiar, experimentar*}; **try**: {***try***, *seek, attempt, essay, assay*}; **apostar**: {*apostar, arriscar, jogar, pôr*}; **gamble**: {***gamble***, *chance, risk, hazard, take chances, adventure, run a risk, take a chance*}; **arriscar**: {*arriscar, aventurar, malparar*}; **risk**: {***risk***, *put on the line, lay on the line*}.

4 FINAL REMARKS

In sum, this paper described the design and content of the current WN.Br database, the procedures and tools for coding synsets, co-text sentences, glosses, language-independent conceptual-semantic relations, and conceptual-semantic equivalence EQ-RELATIONS. The overall procedures for constructing wordnets presented in this paper, though not resorting to reusing existing resources, a current tendency in the field [26], devised a reliable, an efficient, and an automatic way of inheriting WN.Pr´s internal relations in the task of constructing wordnets to other languages.

On the way, besides the specification of the other language-independent conceptual-semantic relations for the verb synsets, it is the encoding of (a) a gloss for each synset of nouns; (b) a co-text sentence for each noun; (c) the mapping of the WN.Br noun synsets to its equivalent ILI-records by means of the following equivalence relations EQ-SYNONYM, EQ-NEAR-SYNONYM, EQ-HAS-HYPERONYM, and EQ-HAS-HYPONYM, and (d) the automatic inheritance from WN.Pr of the relevant conceptual-semantic relations of hyponymy/hypernymy, co-hyponymy, and meronymy/holonymy relations for nouns.

ACKNOWLEDGEMENTS

This project was supported in part by contract 552057/01, with funding provided by The National Council for Scientific and Technological Development (*CNPq*), Brazil; in part by grant 2003/03623-7 from The State of São Paulo Research Foundation (FAPESP), Brazil. My thanks go to all my linguistics students at *CELiC* and the *NILC* developers for their invaluable help in constructing the WN.Br core database. Special thanks to *CAPES*, the *PPG Linguística e Língua Portuguesa, FCL-UNESP/Araraquara*, and the *PROPG-UNESP*.

REFERENCES

1. Palmer, M. (ed.): Miltilingual resources – Chapter 1. In: Eduard Hovy, Nancy Ide, Robert Frederking, Joseph Mariani, and Antonio Zampolli (eds.): Linguistica Computazionale, Vol. XIV-XV (2001)
2. Hanks, P.: Lexicography. In: The Oxford Handbook of Computational Linguistics, R. Mitkov (ed.), Oxford, Oxford University Press (2003)

3. Di Felippo, A., Pardo, T.A.S., Aluísio, S.M. Proposta de uma metodologia para a identificação dos argumentos dos adjetivos de valência 1 da língua portuguesa a partir de córpus. In: Carderno de Resumos do V Encontro de Corpora, São Carlos, São Paulo (2005) 20-21
4. Handke, J.: The structure of the Lexicon: human versus machine. Berlin: Mouton de Gruyter (1995)
5. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge (1998)
6. Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, M.A., Peters, W.: The Linguistic Design of the EuroWordNet Database. Computers and the Humanities, Vol. 32 (1998) 91-115
7. Gonçalo, J., Verdejo, F., Peters, C., Calzolari, N.: Applying EuroWordNet to Cross-Language Text Retrieval. Computers and the Humanities, Vol. 32 (1998) 185-207
8. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities, Vol. 32(2,3)(1998) 73-89
9. Peters, W., Vossen, P., Díez-Orzas, P., Adriaens, G.: Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index. Computers and the Humanities, Vol. 32 (1998) 221-251
10. Dias-da-Silva, B.C., Oliveira, M.F.; Moraes, H.R. Groundwork for the development of the Brazilian Portuguese Wordnet. Advances in natural language processing. Berlin: Springer-Verlag (2002)189-196
11. Dias-da-Silva. B.C.; Moraes, H.R. A construção de thesaurus eletrônico para o português do Brasil. Alfa. São Paulo: Editora Unesp, Vol. 47(2) (2003) 101-115
12. Dias-da-Silva, B.C.: Human language technology research and the development of the Brazilian Portuguese wordnet. In: Proceedings of the 17th International Congress of Linguists – Prague, E. Hajičová, A. Kotěšovcová, J. Mírovský, ed., Matfyzpress, MFF UK (2003) 1-12
13. Hayes-Roth, F.: Expert Systems. In: Encyclopedia of Artificial Intelligence, E. Shapiro (ed.), Wiley, New York (1990) 287-298
14. Durkin, J.: Expert Systems: Design and Development. Prentice Hall International, London (1994)
15. Dias-da-Silva, B. C.: Bridging the Gap Between Linguistic Theory and Natural Language Processing. In: 16th International Congress of Linguists – Paris, B. Caron, ed., Pergamon-Elsevier Science, Oxford (1998) 1-10
16. Rodríguez, H, Climent, S., Vossen, P., Bloksma, L., Peters, W. Alonge, A., Bertagna, F., Roventini, A.: The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top-Ontology. Computers and the Humanities, Vol. 32 (1998) 117-152
17. Miller,G.A.:Dictionaries in the Mind. Language and Cognitive Processes,Vol.1(1986)171-185
18. Miller, G.A., Fellbaum, C.: Semantic Networks of English. Cognition 41 (1991) 197-229

19. Ferreira, A. B. H.: Dicionário Aurélio Eletrônico Século XXI. Lexicon, São Paulo, CD-ROM (1999)
20. Weiszflog, W. (ed.): Michaelis Português – Moderno Dicionário da Língua Portuguesa. DTS Software Brasil Ltda, São Paulo, CD-ROM (1998)
21. Barbosa, O.: Grande Dicionário de Sinônimos e Antônimos. Ediouro, Rio de Janeiro, 550 p. (1999)
22. Nascentes, A.: Dicionário de Sinônimos. Nova Fronteira, Rio de Janeiro (1981)
23. Borba, F.S. (coord.): Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil. Editora da Unesp, São Paulo, 600 p. (1990)
24. Borba, F.S.: Dicionário de usos do português do Brasil. São Paulo: Ed. da UNESP (2002)
25. Houaiss, A.: Dicionário Eletrônico Houaiss da Língua Portuguesa. FL Gama Design Ltda., Rio de Janeiro CD-ROM (2001)
26. Rigau, G., Eneko, A.: Semi-automatic methods for WordNet construction. In: 1st Intenational WordNet Conference Tutorial, Mysore, India (2002)

BENTO CARLOS DIAS-DA-SILVA
CENTRO DE ESTUDOS LINGÜÍSTICOS
E COMPUTACIONAIS DA LINGUAGEM - CELIC1,
FACULDADE DE CIÊNCIAS E LETRAS,
UNIVERSIDADE ESTADUAL PAULISTA (UNESP)
CAIXA POSTAL 174 – 14.800-901, ARARAQUARA, SP, BRAZIL
E-MAIL: <BENTO@FCLAR.UNESP.BR>