The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portuguese WordNet Database

[°]Bento Carlos Dias-da-Silva, [°]Ariani Di Felippo, [+]Maria das Graças Volpe Nunes
[°]CELiC - Centro de Estudos Lingüísticos e Computacionais da Linguagem
Faculdade de Ciências e Letras – Universidade Estadual Paulista (UNESP)
CP 174 – 14.800-901, Araraquara, SP, Brazil
[°+]NILC - Núcleo Interinstitucional de Lingüística Computacional
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
CP 668 – 13.560-970, São Carlos, SP, Brazil
bento@fclar.unesp.br, arianidf@uol.com.br, gracan@icmc.usp.br

Abstract

Princeton WordNet (WN.Pr) lexical database has motivated efficient compilations of bulky relational lexicons since its inception in the 1980´s. The EuroWordNet project, the first multilingual initiative built upon WN.Pr, opened up ways of building individual wordnets, and interrelating them by means of the so-called Inter-Lingual-Index, an unstructured list of the WN.Pr synsets. Other important initiative, relying on a slightly different method of building multilingual wordnets, is the MultiWordNet project, where the key strategy is "building language specific wordnets keeping as much as possible of the semantic relations available" in the WN.Pr. This paper, in particular, stresses that the additional advantage of using WN.Pr lexical database as a resource for building wordnets for other languages is to explore possibilities of implementing an automatic procedure to map the WN.Pr conceptual relations as hyponymy, co-hyponymy, troponymy, meronymy, cause, and entailment onto the lexical database of the wordnet under construction, a viable possibility, for those are language-independent relations that hold between lexicalized concepts, not between lexical units. Accordingly, combining methods from both initiatives, this paper presents the ongoing implementation of the WN.Br lexical database and the aforementioned automation procedure illustrated with a sample of the automatic encoding of the hyponymy and co-hyponymy relations.

1. Introduction

Natural Language Processing (NLP) initiatives to devise methods for developing computational lexicons either manually from scratch or (semi-)automatically from machine-readable dictionaries have attested that coding lexicons for NLP applications is a time-consuming, prone to flaws task (Palmer et al., 2001; Hanks, 2003; Matsumoto, 2003). The core of the problem is the amount, the variety, and the complexity of specialized and interrelated information that lexicon developers have to cope with and to encode in the lexical database: graphemic, morphological, syntactic, semantic, and even illocutionary bits of information, among others (Handke, 1995).

But, on the one hand, there have been important initiatives to minimize the burden of the task and to develop strategies and standards for building robust and corpus-based lexicons (cf. Calzolari, McNaught & Zampolli, 1996; Zampolli, 1997, Lenci et al., 2000), acquiring lexical information from MRD and corpora (Matsumoto & Utsuro, 2000), and establishing the necessary "virtuous circle" model between lexicons and corpora (Calzolari, 2004, p. 102). On the other, a successful psycholinguistic experiment, the Princeton WordNet (WN.Pr) lexical database, a taxonomic thesaurus, has motivated efficient compilations of bulky relational lexicons since its inception in the 1980´s (Miller & Fellbaum, 1991; Fellbaum, 1998).

The EuroWordNet relational lexical database (Vossen, 1998) is the first multilingual initiative built upon WN.Pr and consists of a collection of individual wordnets interrelated by means of the so-called Inter-Lingual-Index (ILI), an unstructured list of the WN.Pr synsets[1].

Other initiative, relying on a slightly different method of building multilingual wordnets, is the MultiWordNet (MWN) project. Pianta, Bentivogli & Girardi (2001, p. 294) argue that the MWN model allows the implementation of automatic procedures to speed up both the construction of the synsets in the target language and the detection of divergences between WN.Pr and the wordnet being built. The key strategy is "building language specific wordnets keeping as much as possible of the semantic relations available" in the WN.Pr (Bentivogli, Pianta & Pianesi, 2000, p.663).

The additional advantage of using WN.Pr lexical database as a resource for building wordnets for other languages, and the one demonstrated in this paper, is to explore possibilities of implementing an automatic procedure to map the WN.Pr hierarchical relations (hyponymy, co-hyponymy, troponymy, meronymy, cause, and entailment) onto the lexical database of the wordnet under construction. It should be stressed that that is, in fact, a viable possibility, for those are language-independent relations that hold between lexicalized concepts. Accordingly, as those relations do not hold between word forms, in wordnets they are appropriately specified between synsets, which are formal entities that represent lexicalized concepts, which legitimately might be lexicalized across languages.

Thus, combining methods from both initiatives, this

[1] An ILI-record consists of a WN.Pr (version 2.0) synset, its concept gloss and its ID number.

paper presents the ongoing implementation of the aforementioned automation procedure. In particular, it focuses on the specification of both the hyponymy and co-hyponymy relations between Brazilian Portuguese WordNet (WN.Br) synsets.

The paper is structured as follows: section 2 sketches out the WN.Br project and the underlying structure of the WN.Br lexical database under construction; section 3 discusses the project alignment issues; section 4 illustrates both the procedure of manual encoding of the <EQ SYNONYM> cross-lingual relations and the automatic specification of the hyponymy and co-hyponymy relations; section 5 concludes the paper with the current WN.Br verb database statistics and the ongoing work; sections 6 and 7 contain the Acknowledgements and the References, respectively.

2. The WN.Br Project

Based on Expert Systems development, the WN.Br project launched in 2003 (Dias-da-Silva, 2003), applies a three-domain approach methodology to develop the WN.Br lexical database, and assumes 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. As such, it must be "mined", "molded", and "assembled" into a computer-tractable system (Durkin, 1994).

2.1 The Three-Domain Philosophy

Accordingly, the process of building the WN.Br lexical database is developed in the following three complementary domains: (a) *the linguistic-related domain*, in which the lexical resources (dictionaries and text corpora), the wordnet lexical-conceptual relations, and the "Base Concepts" and the "Top Ontology" (Vossen, 2003), i.e. the “natural language ontology” of concepts, are mined; (b) *the representational domain*, in which the overall information selected and organized in the preceding domain is molded into computer-tractable representations, e.g. the *synset* (Miller, 1986) – a set of words built on the basis of the notion of synonymy in context, i.e. word interchangeability in some context –, the *lexical matrix* (Miller & Fellbaum, 1991) – associations of sets of word forms and the concepts they lexicalize –, and the wordnet "lexical database" itself (Fellbaum, 1998); (c) *the computational domain*, in which the computer-tractable representations are assembled by means of the WN.Br Editor.

2.2 The WN.Br Underlying Structure

The underlying structure of the WN.Br lexical database shown in Fig.1 is made up of two lists: the List of Entries (LE), i.e. the list of lexical units (arranged in alphabetical order) pertaining to one of the following four syntactic categories: verb, noun, adjective, or adverb; and the List of Synsets (LS), i.e. the collection of the synsets formed from those lexical units. Each lexical unit in a synset is necessarily an element of the LE and is specified for its particular Sense Description Vector (SDV). Each SDV has three pointers: the synonymy pointer, which identifies a particular synset in the LS; the antonymy pointer, which identifies a particular antonym synset in the LS; and the sense pointer, which identifies a particular sense number in the SDV. Each synset is also linked to its concept gloss via the concept gloss link, and each lexical unit is linked to its co-text sentence via the co-text sentence link.

By means of the WN.Br Editor the linguist (a) inserts, consults, modifies, and saves lexical unit types and synsets; (b) inserts co-text sentences, extracted from corpora, for each lexical unit; (c) writes a concept gloss for each synset, and (d) generates synsets lists by syntactic category, by number of elements, by their degree of homonymy and polysemy, and by co-text sentences.

3. Conceptual Alignment Issues

The WN.Br work in progress is the linking of its verb synsets to the equivalent ones in the WN.Pr lexical database by the aforementioned <EQ RELATIONS>. Such a conceptual alignment permits not only the linguistic investigation of differences and similarities in the lexicalization processes between Brazilian Portuguese and American English but also two sorts of mismatches described by Peters et al. (1998): the linguistic mismatches (lexical gaps[2], due largely to cultural gaps, pragmatic differences, and morphological mismatches; over-differentiation or under-differentiation of senses; and fuzzy-matching between synsets) and technical mismatches (mistakes in the choice of inter-lingual equivalence links or in the encoding of language-independent relations across wordnets).

3.1 The WN.Br Lexical Database Structure

The WN.Br Editor, a Windows®-based , besides aiding the linguist in the manual encoding of both the WN.Br synsets and the cross-lingual equivalence relations between synsets, the so called <EQ RELATIONS>[3] (Vossen et al., 1998; Peters et al., 1998), makes it now possible to encode the WN.Br language-internal relations of hyponymy, co-hyponymy, troponymy, cause, and entailment automatically by inheriting them from the WN.Pr lexical database.

To cope with these tasks, the original WN.Br Editor (Dias-da-Silva, 2003) was enhanced to house the three interconnecting modules described in the next section. Accordingly, the original WN.Br lexical database underlying structure shown in Fig.1 was extended to encode the EQ-RELATIONS (see Fig.2).

Thus, each synset structure was augmented with an additional vector to identify both the wordnet standard language-independent conceptual relations (e.g. hyponymy and co-hyponymy) and the cross-lingual <EQ RELATIONS> between synsets of the two wordnets. This new vector enriched the WN.Br database structure with the

[2] Bentivoglio & Pianta (2000) propose a procedure for identifying lexical gaps semi-automatically.

[3] <EQ SYNONYM>, <EQ NEAR SYNONYM>, <EQ HAS HYPONYM>, <EQ CAUSES>, and <EQ IS SUBVENT OF>.

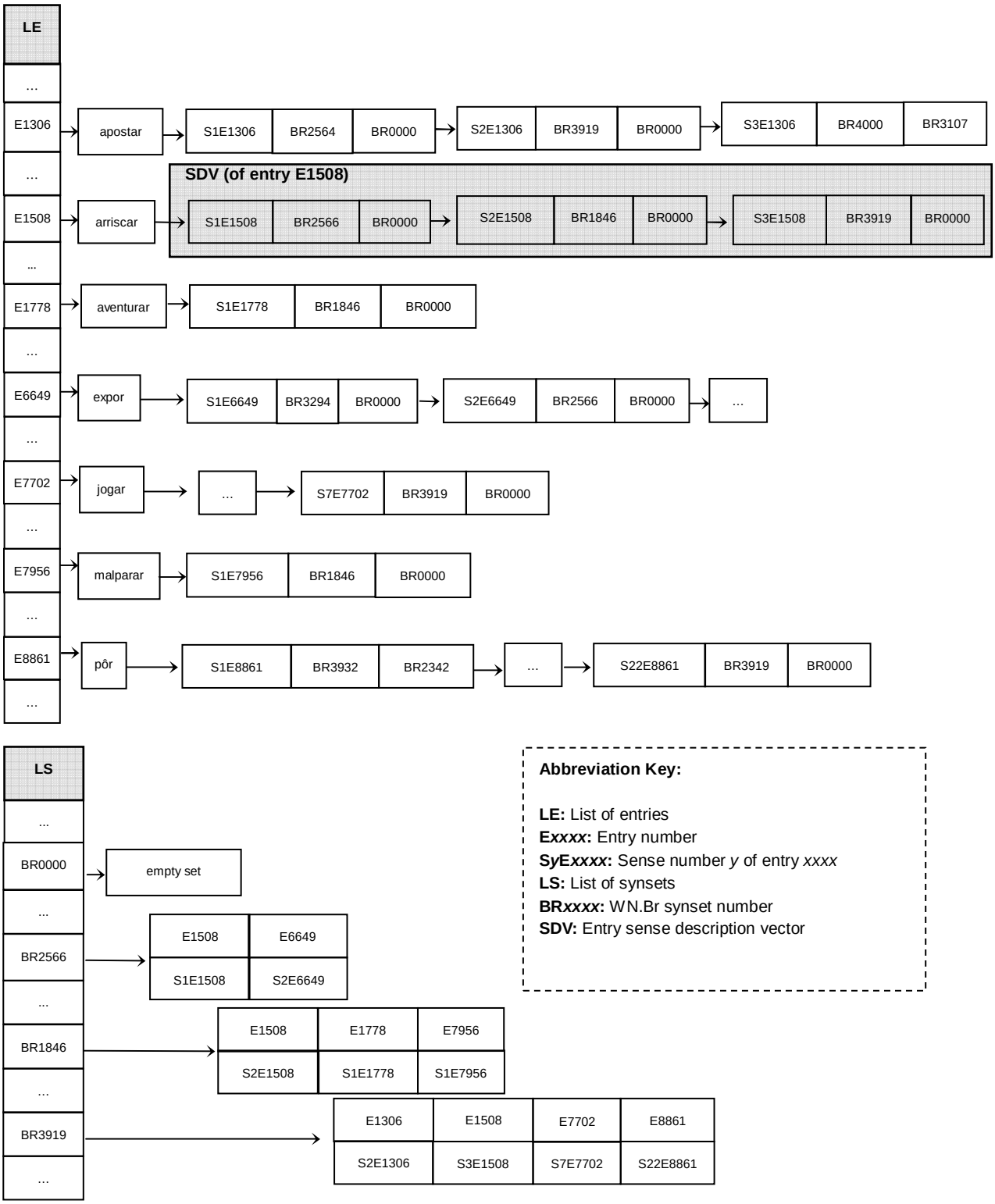

Figure 1: The WN.Br database underlying structure generated by the editing tool.

following cross-linguistic information:

- a 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 concept gloss, e.g. Expose to a chance of loss or damage,
- the English co-text sentence, e.g. "Why risk your life?",
- EQ-RELATIONS, e.g. EQ-SYNONYM ({arriscar, expor}, {risk, put on the line, lay on the line}).

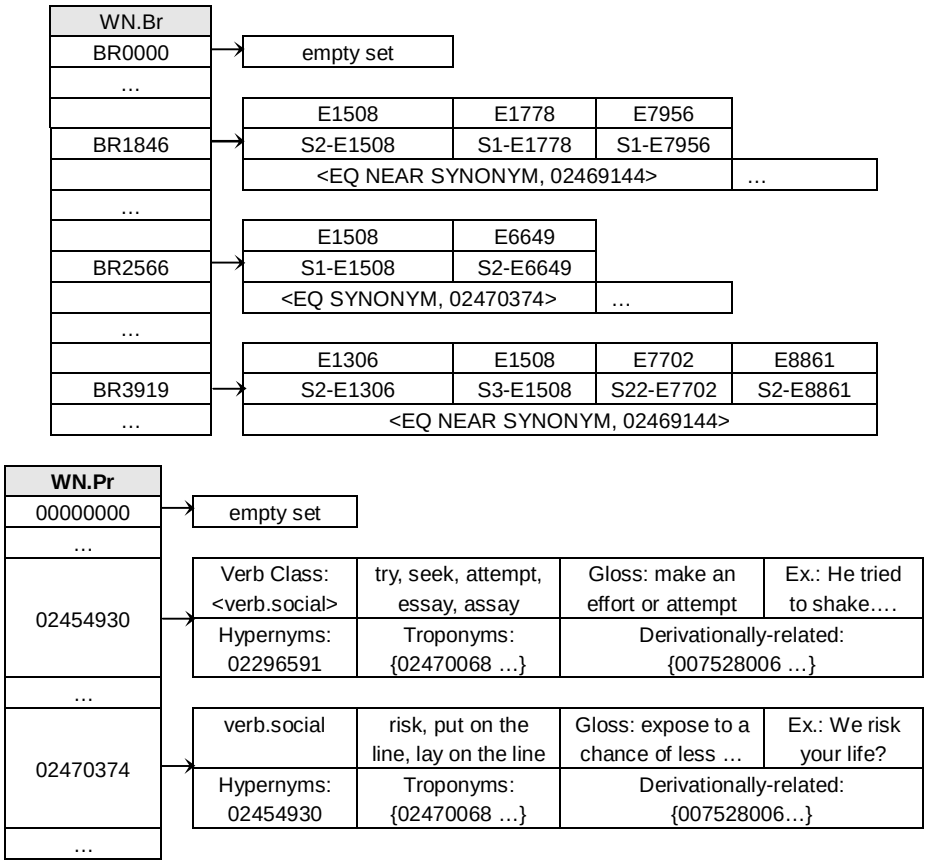

Figure 2: The augmented WN.Br database underlying structure with <EQ RELATIONS>.

4. Alignment of WordNets and Automatic Encoding of Language-Internal Relations

A brief example will illustrate both the manual alignment procedure to map WN.Br synsets onto WN.Pr synsets (section 4.1) and the automatic specification of the language-internal relations of hyponymy and co-hyponymy (section 4.2).

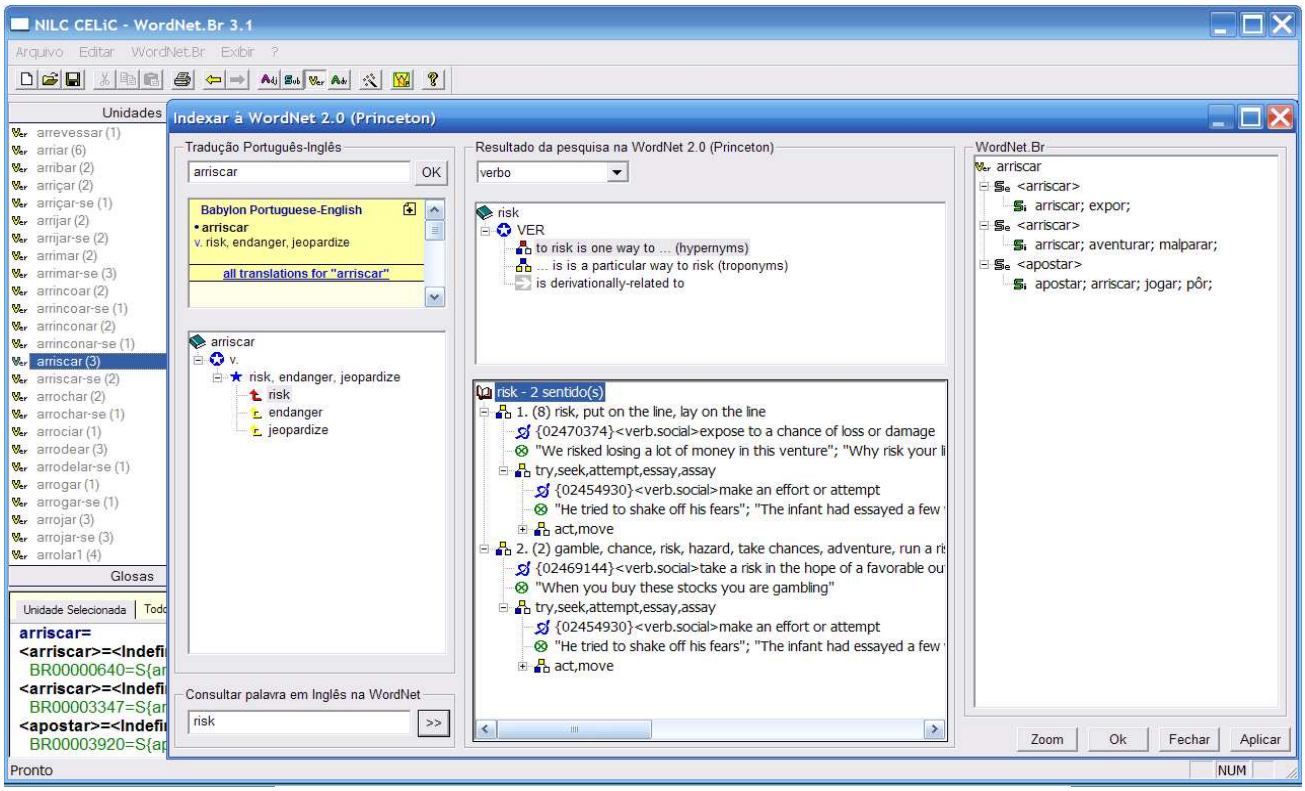

Figure 3a: The WN.Br Editor three-column window.

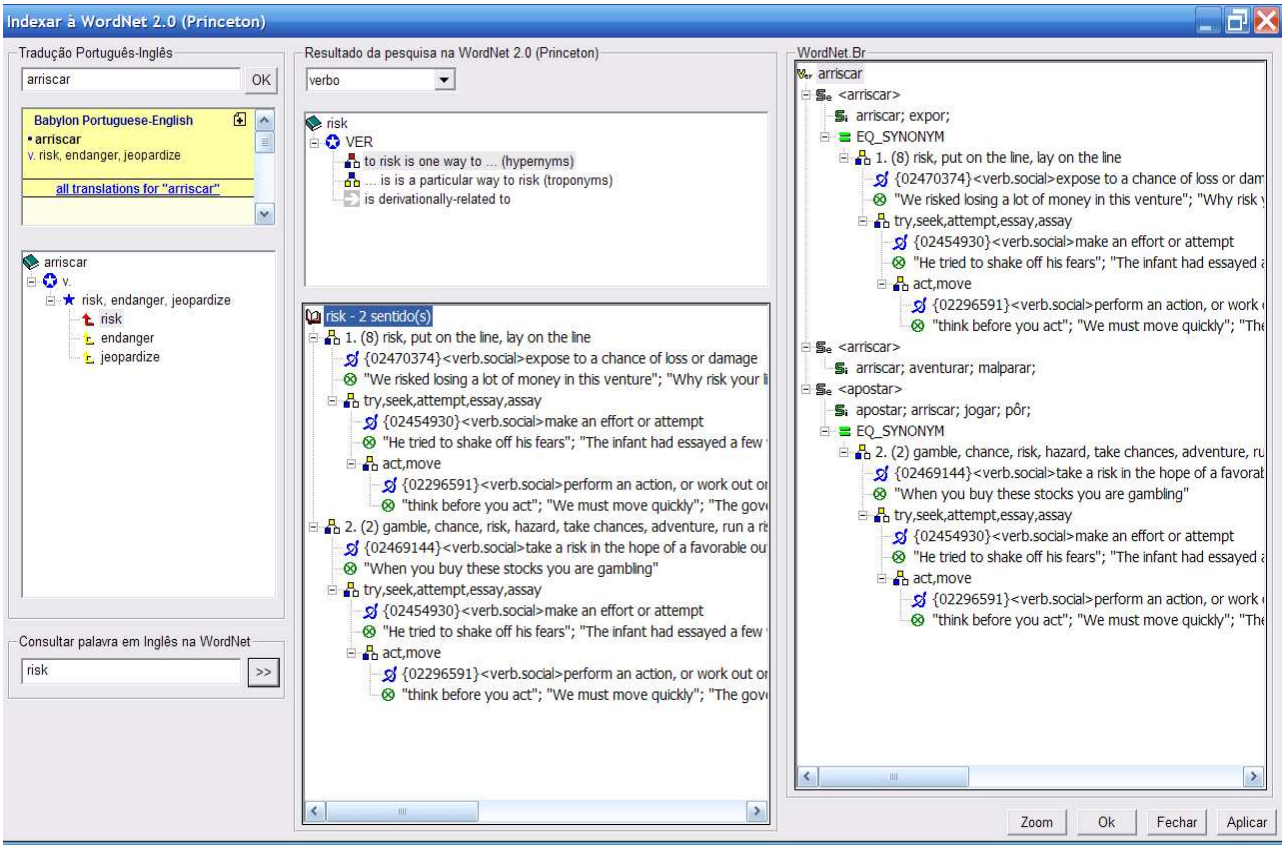

Figure 3b: A sample of two manual alignments.

4.1 The Manual Alignment

The linguist starts off the procedure by selecting a verb in the WN.Br Editor list (e.g. *arriscar*). As shown in Fig.3a, the editor three-column alignment window pops up. The left column displays the bilingual search results of the editor dictionary look-up tool[4]: *arriscar* ↔ *risk*, endanger, jeopardize. The right column displays the following three WN.Br synsets: {*arriscar*, *expor*}; {*arriscar*, *aventurar*, *malparar*}; {*apostar*, *arriscar*, *jogar*, *pôr*}.The center column, in turn, displays the following two WN.Pr synsets, which contain the English verb *risk* selected by the linguist from the search results in the left column: {*risk*, *put on the line*, *lay on the line*}; {*gamble*, *chance*, *risk*, *hazard*, *take chances*, *adventure*, *run a risk*, *take a chance*}.[5]

In the next step, the linguist drags and drops the

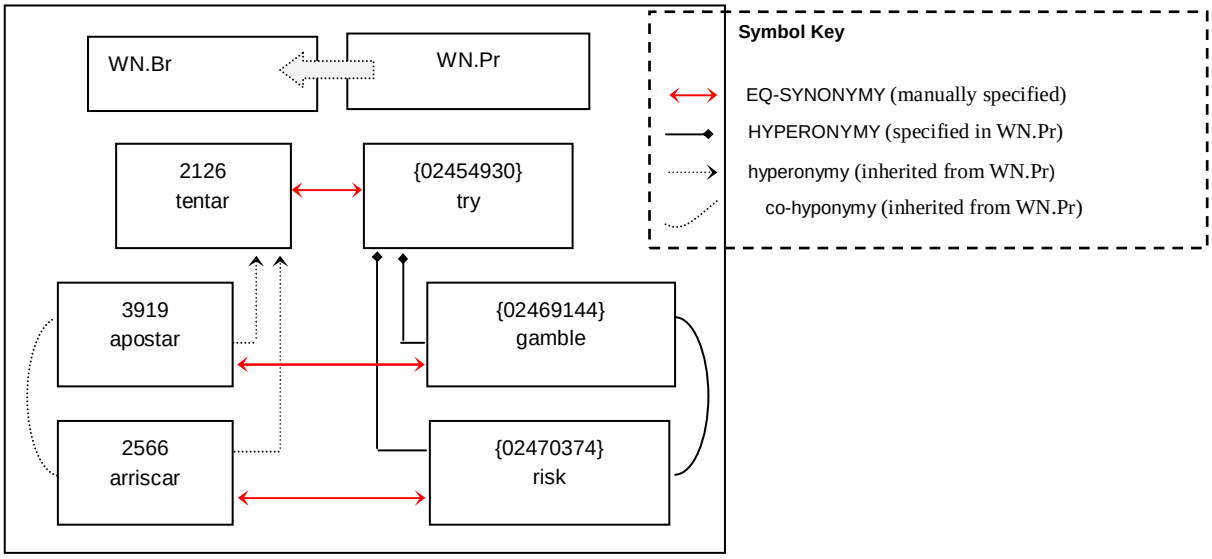

Figure 4: A sample of the automatic encoding.

[4] The editor dictionary look-up tool searches the *Babylon Portuguese-English* dictionary online automatically.

[5] The information in this column is formatted according to the *WordNet TreeWalk* Applet (Beau, 2003).

appropriate WN.Pr synsets from the center column (synsets numbered 02470374 and 02469144) onto the appropriate WN.Br synsets in the right column. The default link label is <EQ SYNONYM>. The resulting alignment is shown in Fig3b.

4.2 The Automatic Encoding

Both the manual and the automatic encoding are illustrated in Fig.4, where red double-headed arrows indicate manual alignments and dashed curve and arrows highlight the automatic encodings of the hyponymy and co-hyponymy relations.

5. Conclusions

Compared to the standard methodologies, which resorts to pre-existing MRDs (Rigau & Eneko, 2002), this paper presented procedures and an original editing tool for encoding both the language-internal wordnet bits of information (synsets, semantic types, glosses, and lexical-conceptual hierarchical relations) and the cross-lingual relations. The latter, the so called <EQ RELATIONS>, has made it possible to connect the two wordnets and to devise a procedure that allows for the automatic encoding of the WN.Br lexical internal hierarchical relations.

In these years of research, the WN.Br lexical database has circa 11,000 verbs (4,000 synsets), 17,000 nouns (8,000 synsets), 15,000 adjectives (6,000 synsets), and 1,000 adverbs (500 synsets) (Dias-da-Silva 2003). Its current 18,500 synsets (44.000 word types) were handcrafted by a team of three linguists, who reused, merged, and tuned synonym and antonym information registered in seven bulky dictionaries, and mined/filtered relevant lexical information from Brazilian Portuguese texts in corpora and in the web to further specify and check these and the other wordnet-related conceptual relations.

On the way, it is the manual encoding of (a) a co-text sentence for each verb, (b) a concept gloss for each synset of verbs; (c) the mapping of each WN.Br verb synset onto its equivalent ILI-record by means of one the aforementioned <EQ RELATIONS>, and (d) the automatic encoding of the aforementioned language-internal relations. Circa three thousand <EQ SYNONYM> relations between WN.Br and WN.Pr synsets have already been encoded.

6. Acknowledgements

My thanks go to the LREC2008 referees, who helped make this paper better, and to Ricardo Hasegawa (NILC-USP computer scientist), who has been working on the WN.Br Editor for all these years. This project is supported in part by Funding 552057/01 from The Brazilian National Council for Scientific and Technological Development (CNPq), and in part by Grants 2006/04447-6 and 2007/01514-7 from The State of São Paulo Research Foundation (FAPESP).

7. References

Beau, B. (2003). WordNet TreeWalk Applet. SourceForge Net. http://wntw.sourceforge.net.

Bentivogli, L., Pianta, E. (2000). Looking for lexical gaps. In *Proceedings of the Ninth EURALEX International Congress*, Stuttgart, Germany, August 8-12, 2000, pp. 663-669.

Bentivogli, L., Pianta, E., Pianesi, F. (2000). Coping with lexical gaps when building aligned multilingual wordnets. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May- 2 June 2000, pp. 993-997.

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.

Calzolari, N., McNaught, J., Zampolli, A. (1996). *EAGLES Final Report: Editor´s Introduction*. Pisa: Eagles.

Dias-da-Silva, B.C. (1998). Bridging the Gap Between Linguistic Theory and Natural Language Processing. In *Proceedings of the Sixteenth International Congress of Linguists*, Paris: Pergamon-Elsevier Science, pp. 1-10.

Dias-da-Silva, B.C. (2003). Human Language Technology Research and the Development of the Brazilian Portuguese WordNet. In *Proceedings of the Seventeenth International Congress of Linguists*, Prague: Matfyzpress, pp. 1-12.

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

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

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

Hanks, P. (2003). Lexicography. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, pp. 48-69.

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.

Matsumoto, Y. (2003). Lexical knowledge acquisition. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, pp. 395-413.

Matsumoto, Y., Utsuro, T. (2000). Lexicography. In R. Dale, H. Moisl & H. Somers (Eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, pp. 563-610.

Miller, G.A. (1986). Dictionaries in the mind. *Language and Cognitive Processes*, 1, pp. 171-185.

Miller, G.A., Fellbaum, C. (1991). Semantic networks of English. *Cognition*, 41(1), pp. 197-229.

Palmer, M., Calzolari, N., Choukri, K., Fellbaum, C., Hovy, E., Ide, N. (2001). Multilingual resources. *Linguistica Computazionale*, 14-15, pp. 1-33.

Peters, W., Vossen, P., Díez-Orzas, P., Adriaens, G. (1998). Cross-linguistic alignment of WordNets with an Inter-Lingual-Index. *Computers and the Humanities*, 32

(2-3), pp. 221-251.

Pianta, E., Bentivogli, L., Girardi,C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January 21-25, 2002, pp. 293-302

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

Vossen, P. (2003). Ontologies. In: R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, pp. 464-482.

Vossen, P., Bloksma, L., Alonge, A., Marinai, E., Peters, C., Castellon, I., Marti, A., Rigau, G. (1998). Compatibility and interpretation of relations in EuroWordNet. *Computers and the Humanities*, 32 (2-3), pp. 153-184.

Rigau, G., Eneko, A. (2002). Semi-automatic methods for WordNet construction. In *First Intenational WordNet Conference Tutorial*, Mysore, India.

Zampolli A. (1997). The PAROLE project in the general context of the European actions for Language Resources. In *Telri Proceedings of the Second European Seminar: Language Applications for a Multilingual Europe*. Manheim/Kaunas: IDS/VDU, pp. 185-210.