Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations

Bento C. Dias-da-Silva[1], Ariani Di Felippo[2], Ricardo Hasegawa[3]

Centro de Estudos Lingüísticos e Computacionais da Linguagem - CELiC[1,2]
Faculdade de Ciências e Letras – Universidade Estadual Paulista (UNESP)
Caixa Postal 174 – 14.800-901, Araraquara, SP, Brazil
Núcleo Interinstitucional de Lingüística Computacional – NILC[1,2,3]
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970, São Carlos, SP, Brazil
bento@fclar.unesp.br arianidf@uol.com.br rh@icmc.usp.br

Abstract. This paper presents the overall methodology that has been used to encode both the Brazilian Portuguese WordNet (WordNet.Br) standard language-independent conceptual-semantic relations (hyponymy, co-hyponymy, meronymy, cause, and entailment) and the so-called cross-lingual conceptual-semantic relations between different wordnets. Accordingly, after contextualizing the project and outlining the current lexical database structure and statistics, it describes the WordNet.Br editing GUI that was designed to aid the linguist in carrying out the tasks of building synsets, selecting sample sentences from corpora, writing synset concept glosses, and encoding both language-independent conceptual-semantic relations and cross-lingual conceptual-semantic relations between WordNet.Br and Princeton WordNet.

1 Introduction

On the one hand, NLP community initiatives to devise methods for developing computational lexicons either from scratch or (semi-)automatically from machine readable dictionaries (MRD) have attested how time-consuming and prone to flaws is to code lexicons for NLP applications [1], [2], [3]. In fact, the bulk of the problem has to do with the amount, the variety, and the complexity of specialized and interrelated information lexicon developers have to cope with and to encode in the database: phonetic/graphemic, morphological, syntactic, semantic, and even illocutionary bits of information [4].

On the other hand, Princeton WordNet (PWN), a successful psycholinguistic experiment, has set the pattern for compiling bulky relational lexicons since its inception in the 1980´s. PWN is basically an on-line relational semantic database combining the design of both a dictionary and a thesaurus. Like 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]. Like a thesaurus, words are grouped in terms of concepts, which are, in turn, represented in terms of synonym sets (*synsets*), i.e. sets of words of the same syntactic category that lexicalizes the same concept. Its web structure makes it possible for the user to find a word meaning not only in terms of other words of the same synset but also in terms of its relations to other words in other synsets as well. Despite the fact that PWN is essentially a particular semantic network, its sought-after NLP applications have been discussed by the research community [6], [7].

Structured along the same lines as PWN, wordnets of other languages are under development. The outstanding multilingual initiative is EuroWordNet (EWN) [8], a multilingual database containing monolingual wordnets and equivalence relations for each language synset to the closest concept from the so-called Inter-Lingual-Index (ILI)[1], which enables cross-lingual comparison of words, concept lexicalizations, and meaning relations in different wordnets [9].

Launched in 2003, the WordNet.Br (Brazilian Portuguese WordNet, WBR) extends the Brazilian Portuguese Thesaurus [10], [11]. It 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 of hypernymy[2], hyponymy[3], meronymy (part-whole relation), entailment[4] and cause[5] between synsets.

This paper describes the three aforementioned encoding strategies. Section 2 briefly depicts the current WBR database and its editing GUI (Graphical User Interface), designed to aid the linguist in carrying out the tasks of building synsets, selecting co-text sentences from corpora, and writing synset concept glosses. Section 3 addresses issues of cross-linguistic alignment of wordnets by means of the ILI and desribes the conceptual-semantic alignment strategy adopted to link WBR to PWN. Section 4 outlines the semi-automatic strategy for mapping the PWN verb hyponymy and co-hyponymy relations on to the WBR verb database. Section 5 concludes with some further work.

2 The Current WordNet.Br Lexical Database

After three years of research, the current WBR database presents the following figures: 11,000 verbs (4,000 synsets), 17,000 nouns (8,000 synsets), 15,000 adjectives (6,000 synsets), and 1,000 adverbs (500 synsets), amounting to 44,000 words and 18,500 synsets [12].

Assuming a compromise between Human Language Technology and Linguistics, and based on the Artificial Intelligence notion of Knowledge Representation [13], [14], the project applies a three-domain approach methodology to the development of the database.[6] This approach 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 database core is developed in the following complementary domains: (a) in *the linguistic-related domain*, the lexical resources (dictionaries

---

[1] The ILI is a list made up of each synset of the PWN with its concept gloss (an informal lexicographic definition of the concept evoked by the synset).

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

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

[4] 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

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

[6] This project was supported in part by contract 552057/01, with funding provided by The National Council for Scientific and Technological Development (CNPq); in part by grant 2003/03623-7 from The State of São Paulo Research Foundation (FAPESP).

and text corpora), the lexical and conceptual-semantic relations, and a kind of natural language ontology of concepts ("Base Concepts" and "Top Ontology" [16]) are mined; (b) in *the representational domain*, the overall information selected and organized in the preceding domain is molded into a computer-tractable representation (the "synsets", the "lexical matrix", and the wordnet "lexical database" itself) [5]; (c) in *the computational domain*, the computer-tractable representations are assembled by means of the WordNet.Br editing GUI.

2.1 The Linguistic-related Domain

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

Given the team of three linguists, the unavailability of Brazilian Portuguese MRDs and other computer tractable resources, and a two-year deadline to present large-scale results, the developers, manually, reused, merged, and tuned synonymy and antonymy information registered in five outstanding standard dictionaries of Brazilian Portuguese (BP): [19], [20], [21], [22], and [23, 24].[8] BP texts available in the NILC Corpus[9] and in the web complemented the project reference corpus.

2.2 The Representational Domain

From the logical point of view, the overall structure of the database is made up of two lists: the List of Headwords (LH), the list of words (arranged in alphabetical order), and the List of Synsets (LS), the list of synsets (Fig.1). Each element of a synset (a word form) is necessarily an element of the LH. Each word is specified for its particular Sense Description (SDv) vector. Each SDv is indexed by 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 word form sense number in the SDv. Given such an underlying structure, each synset is linked to its concept gloss via the “concept gloss link”, and each word is linked to its co-text sentence via the “co-text sentence link”.

---

[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 synonyny and antonymy to define word senses. In a way, this choice dictated the strategy to proceed the work alphabetically, instead of working by semantic fields.

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

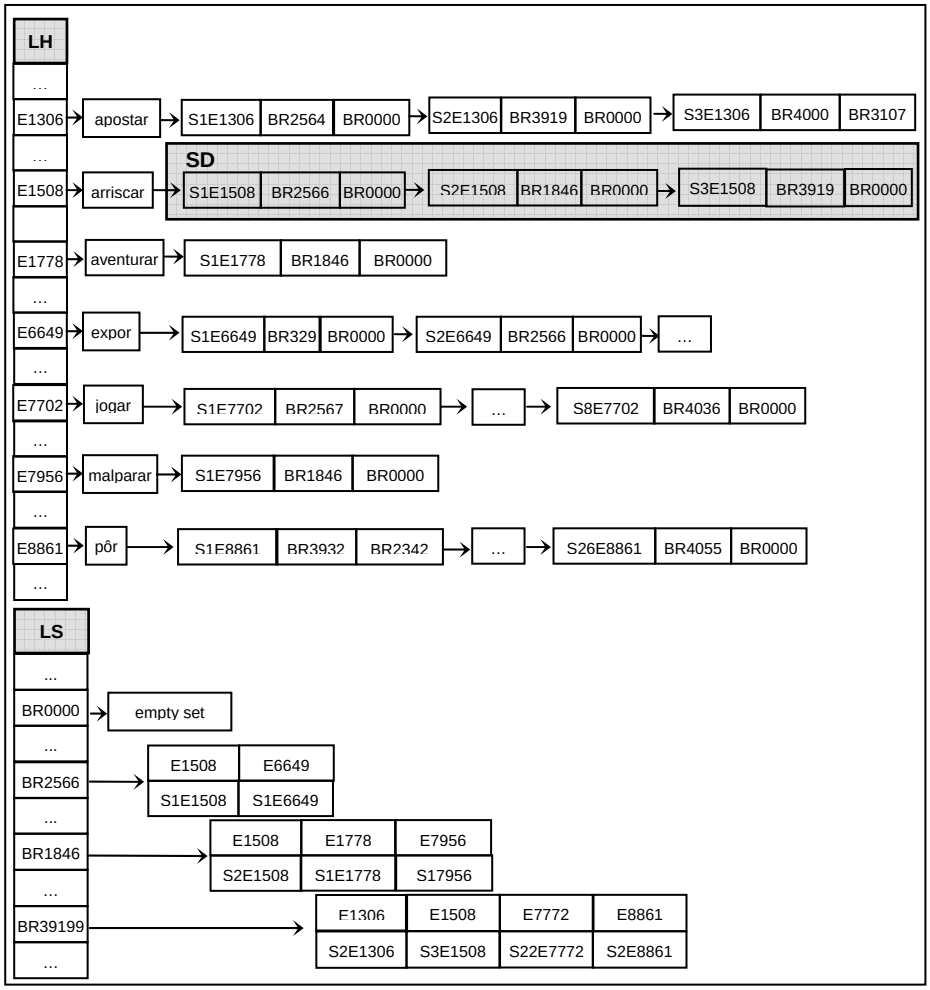

Figure 1. The WordNet.Br underlying structure.

2.3 The Computational Domain

The current WBR editing tool is a Windows®-based GUI. It allows the linguist (a) to create, consult, modify, or save words and synsets; (b) to include co-text sentences for each word; (c) to write a concept gloss for each synset; and (d) to generate different types of synset lists (lists arranged by syntactic category, by number of elements, by the degree of homonymy and polysemy, and by co-text sentence) and different statistics. Its main functionalities include the storage and bookkeeping of the general information of the database. The processes of editing (a) words, and (b) co-text sentences and (c) concept glosses can be better understood by an illustrative example. The first GUI dialogue box in Fig. 2 shows the editor at the moment the

linguist is constructing synsets that contain the verb "lexicalizar" ("to lexicalize"). In the first dialogue box, the linguist selects the appropriate syntactic category and the expected number of senses (i.e. the number of synsets to be constructed); then, s/he clicks on the "Avançar" button ("Next" button). The second dialogue box "Todas as Unidades" field ("All Unities" field) pops up, i.e. the list of all the words already in the database. To construct a synset (or an antonym synset), the linguist picks out the appropriate words from the list and clicks on the "Avançar" button. In the third dialogue box, s/he concludes the synset construction procedure.

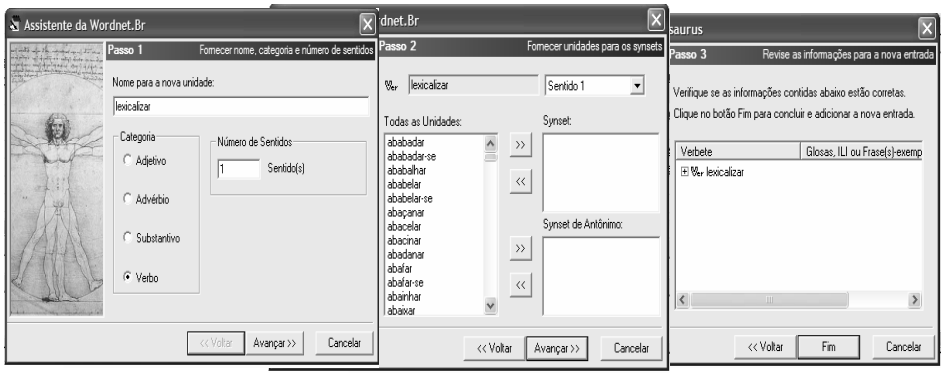

Figure 2. The procedure for encoding synsets.

While words and synsets are inserted through dialogue boxes, the co-text sentences and concept glosses are typed in directly in the editor window (Fig.3). The screen shot to the left illustrates the "Frase(s)-exemplo" field ("Co-text sentence" field) when the linguist clicks on a word. The screen shot to the right illustrates the "Glosa" field ("Gloss" field). Similarly, to type in a concept gloss, the linguist clicks on the synset located in the "Todas as Unidades" field.

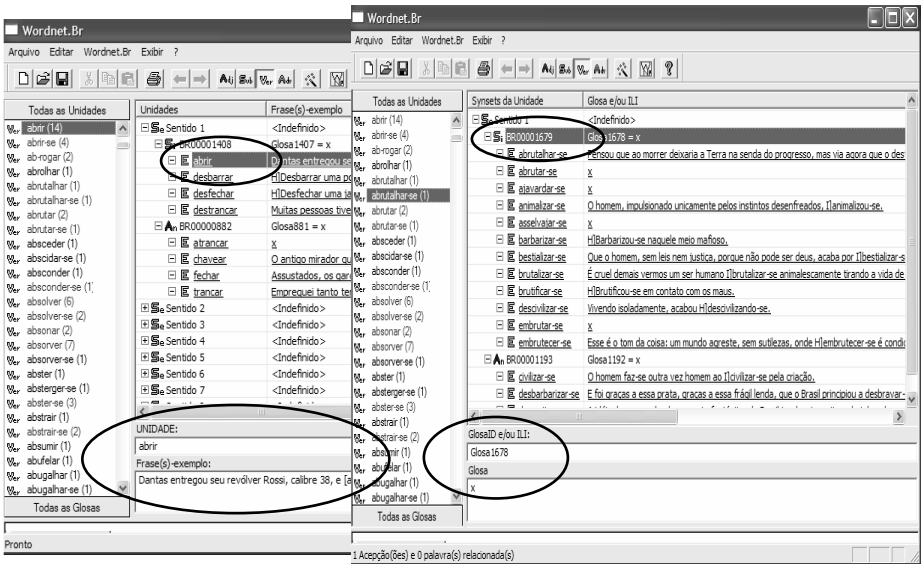

Figure 3. The procedure for encoding co-text sentences and concept glosses.

Currently, the database contains 19,747 co-text sentences selected from the project reference corpus. The following statistics are generated by the editor: 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	Number 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

Number of co-text sentences per synset	Number of synsets
1	18,604
2	521
3	10

3 The Cross-Linguistic Alignment of Wordnets

A rewarding and necessary challenge to the WBR project is to link WBR and PWN (2.0 version) databases. This alignment might permit not only the linguistic investigation of differences and similarities in the lexicalization processes between Brazilian Portuguese and English but also the development of a bilingual lexical database which can be used directly in applications such as cross-language information retrieval involving both languages. Moreover, this bilingual database could generate two types of machine-readable dictionaries: a monolingual Brazilian Portuguese dictionary and a bilingual English-Portuguese dictionary [12]. Furthermore, the possibility of mapping WBR on to PWN might allow the semi-automatic specification of the relevant hierarchical conceptual-semantic relations mentioned in section (1) above.

4 The Alignment Process

The inter-lingual equivalence relations between wordnets are mined in accordance with the types identified by Vossen [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 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) as have been described in Peters [9] are also accounted for during the linking procedures. The salient equivalence relations and cross-lingual possible mismatches are molded into a computer-tractable representation that relies on the unstructured list of the PWN synsets, the aforementioned ILI, conceived of as a kind of interlingua used to link different wordnets.

Specifically, different wordnets are linked by ILI-records[10]. The ILI-record as a linking device has some technical advantages: (a) it is most beneficial with respect to the effort needed for the development, maintenance, future expansion, and reusability of a multilingual wordnet; (b) it avoids the need to develop and maintain a huge and complex semantic structure to incorporate the meanings encoded by each individual wordnet into the multilingual wordnet; (c) it makes less costly for wordnet developers to add new wordnets to the multingual wordnet [9].

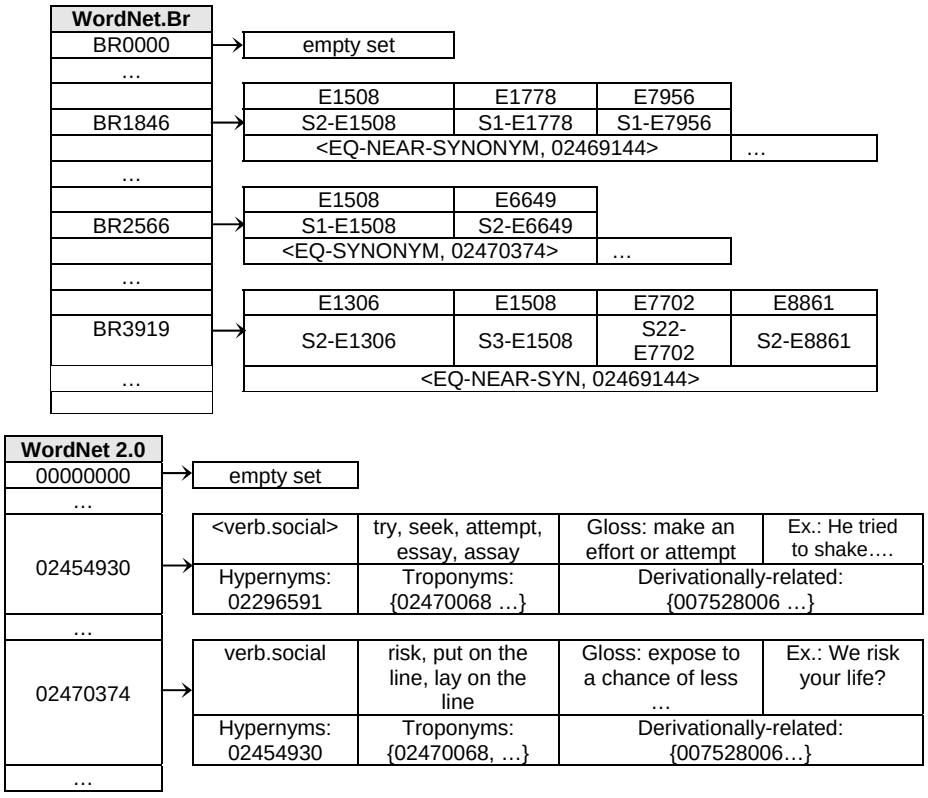

Figure 4. The synset structure augmented with conceptual-semantic EQ-RELATIONS.

To encode the inter-lingual equivalence relations, the overall structure of the database has been further extended as shown in Fig.4. Besides the LH and LS lists and SDv pointers (see 2.2), each synset structure has been augmented with an additional vector to identify both the wordnet standard language-independent concepual-semantic relations (e.g. HYPONYMY, TROPONYMY, CO-HYPONYMY, etc.) and the cross-lingual conceptual-semantic EQ_RELATIONS between synsets of the two wordnets. This new vector enriches the WBR 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}),

[10] An ILI-record is a PWN (version 2.0) synset, its concept gloss and its ID number [9].

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?"), and EQ-RELATIONS (e.g. EQ-SYNONYM relation).

The current WBR editing GUI has three interconnecting modules. Each module, in turn, makes it possible for the linguist to carry out specific tasks during the procedure for linking synsets across the two wordnets: searching the WBR database, the BP-English dictionary, and the web; searching the PWN database automatically; and linking synsets within WBR and across the two wordnets.

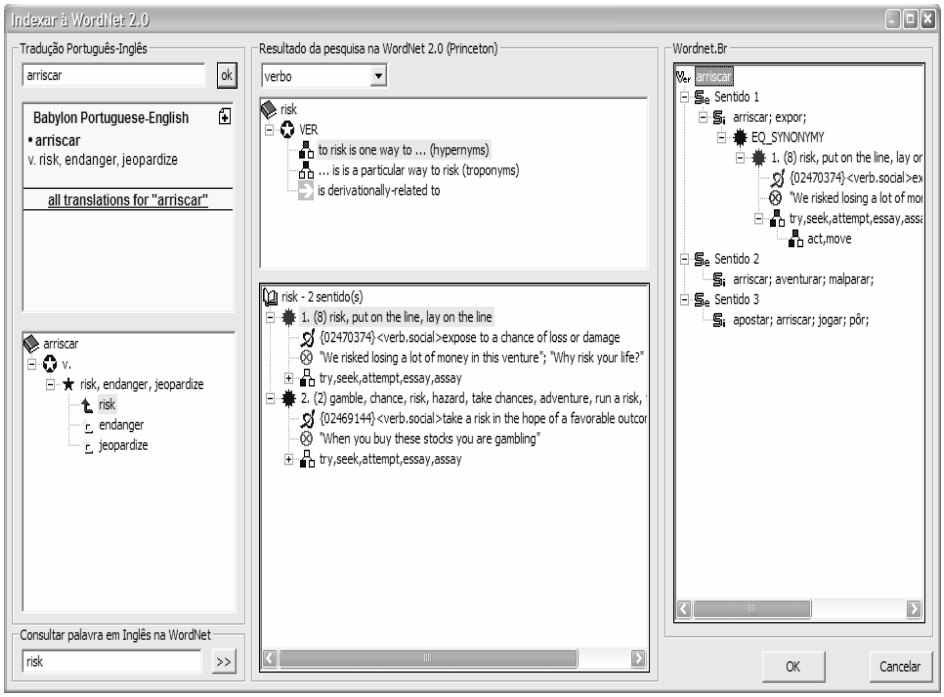

Figure 5. The three-column WordNet.Br GUI.

The linguist starts off the linking procedure by right clicking on a target WBR word. As shown below in Fig. 5, in response to that action the editor displays a three column GUI (the three interconnecting modules), with an online MRD bilingual BP-English dictionary and a web search field at the left, the relevant PWN synsets in the middle, and the WBR synsets that contain the target word to the right. In the first column, (i) the linguist analyzes all possible English words thar are equivalent to the target Brazilian Portuguese word (e.g. the English verbs "risk, endanger, jeopardize" and the BP verb "arriscar"), with recourse to the dictionary and a quick web search;[11] in the middle column, (ii) the linguist analyzes the possible types of equivalence links between the two sets of synsets: the one in the middle column –the sets of

[11] It is also possible to select the appropriate English equivalent (e.g. "risk") to trigger the relevant PWN information in the middle column.

synsets of PWN (e.g. the synsets {risk, put on the line, lay on the line} and {gamble, chance, risk, hazard, take chances, adventure, run a risk, take a chance}– and the one in the column to the right –the WBR synsets that contain the targer word (e.g. the synsets {arriscar, expor}, {arriscar, aventurar, malparar}, and {apostar, arriscar, jogar, pôr}).

5 Conclusion

On the way, it is the encoding of (a) a concept gloss for each synset of verbs; (b) a co-text sentence for each verb; (c) the mapping of the WBR verb 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 the automatic inheritance of PWN´s hipernymy and co-hyponymy relations (See Fig. 6); (d) the conceptual-semantic relations of hypernymy, entailment, and cause between WBR verb synsets.

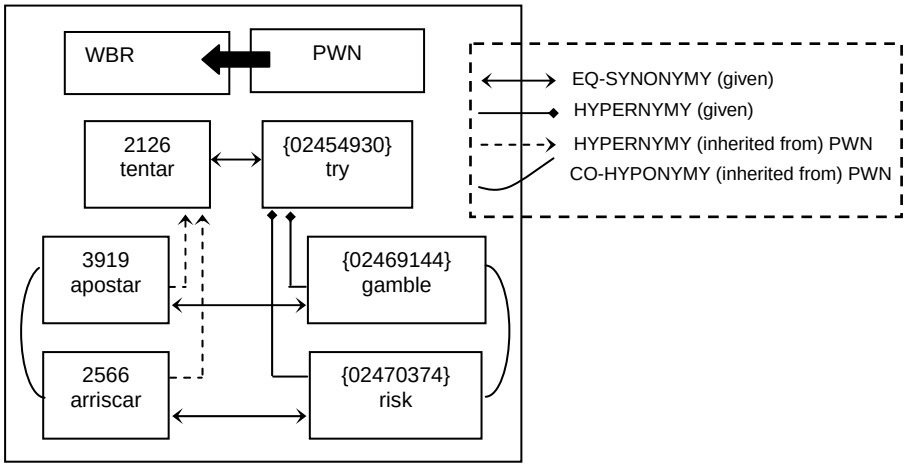

Figure 6. A sample of an automatic encoding of hypernymy and co-hyponymy.

This paper described the overall design and content of the current WBR database, the procedures and tools for encoding synsets, co-text sentences, concept glosses, language-independent conceptual-semantic relations, and conceptual-semantic equivalence relations between WBR and PWN. It should be stressed that the overall procedures described in this paper are efficient and original if compared to the standard methodologies presented by Rigau et al. [26], which resorts to pre-existing MRD lexical resources.

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 International WordNet Conference Tutorial, Mysore, India (2002)