Human Language Technology Research and the Development of the Brazilian Portuguese Wordnet *

Bento Carlos DIAS-DA-SILVA

Faculdade de Ciências e Letras, Universidade Estadual Paulista,
Rodovia Araraquara-Jau Km 1, 14800-901 Araraquara, São Paulo, Brazil
bento@fclar.unesp.br

Abstract: This paper discusses particular linguistic challenges in the task of compiling the Brazilian Portuguese Wordnet, the Wordnet.Br. After setting the scene by overviewing methodological issues, it focuses on the basic steps taken to compile the Wordnet.Br core database: a machine-tractable thesaurus-like lexical database. The discussion is split between three domains: the Linguistic Domain, the Representation Domain, and the Computational Domain.

1. Human Language Technology and Linguistics

It is a fact that an overwhelming growth in Human Language Technology research (henceforth HLT) has taken place since the potential for building computer models of natural language understanding and generation was recognized by the pioneers of machine translation in the early 1950's. As a result, natural language processing (henceforth NLP) has become a discipline in ferment, and gathers researchers with a wide range of backgrounds and interests, emphasizing its diverse aspects, and employing manifold methods and techniques.

Despite the enthusiasm, there have been drawbacks, some of which due to either lack of appreciation for the complexity of natural languages or underspecification of the complexity of the task itself, which reveals a disturbing gap between HLT and Linguistics. Furthermore, Linguistics has either disregarded computational issues altogether or provided the ammunition to deaden the enthusiastic development of NLP technologies.

On the one hand, the HLT challenge is to develop both user-visible NLP applications (e.g., spell and grammar checkers, machine translation systems, information retrieval systems text/speech synthesis, and recognition systems)

* This research is sponsored by CNPq-Brasília and FAPESP-São Paulo, Brazil.

and user-transparent NLP components (e.g., grammars, parsers, tree-banks, lexicons, and lexical resources).

On the other hand, the NLP task is to emulate a particular type of a knowledge processing system where complex linguistic and extra-linguistic pieces of knowledge are formally represented and electronically applied to exploit and to perform a number of linguistic as well as metalinguistic tasks: "check" spelling and grammar, "analyze" morphological and syntactic structures, "understand" and "produce" texts, "translate" words, sentences and texts, "make" and "answer" questions, and "help" linguists themselves develop their own linguistic models (Dias-da-Silva, 1996).

We assume a compromise between HLT and Linguistics and, based on the Artificial Intelligence notion of Knowledge Representation Systems (Hayes-Roth, 1990, Durkin, 1994), propose a the three-domain approach methodology that claims that the linguistic knowledge (i.e., linguistic information) needed to feed NLP systems, like a rare metal, must be mined (the elicitation of the relevant general linguistic information and usage), molded (the computer-tractable representation of that information), and assembled (the computational encoding of the resulting representation into and by means of computer programs). It amounts to saying that the process of designing and implementing NLP systems (i.e., the HLT research itself) should comprise the following iterative and evolutionary phases of analysis in three complementary domains:

- The Linguistic Domain (the mining phase), where the elicitation of the relevant general linguistic information and usage is made;
- The Representational Domain (the molding phase), where the computer-tractable representation of that information is dealt with;
- The Computational Domain (the assembling phase), where the computational encoding of the resulting representation into and by means of computer programs is tackled.

Accordingly, the process of implementing the Wordnet.Br has been split between three complementary domains. This paper, in particular, resorts to the three-domain approach methodology to discuss the initial compilation stage of the Brazilian Portuguese Wordnet (henceforth Wordnet.Br): the task of sorting over 44,000 Brazilian Portuguese words into a machine-tractable thesaurus-like lexical database (henceforth the Wordnet.Br core database),

which is the building block of Wordnet.Br, after Princeton´s WordNet, with capital "N", and EuroWordNet.[1]

In the Linguistic Domain, basic notions of thesaurus and meaning similarity, and strategies for reusing published dictionaries as reference corpus and for mining synonym sets are set up; in the Representation Domain, the representation scheme for lexical meanings and sense relations is established, plus the overall lexical database design; and in the Computational Domain, the editing tool and the statistics of Wordnet.Br core database are sketched.

2. The Linguistic Domain

2.1 The *Thesaurus* Denotations

In what follows, we present a survey of the denotations of the term *thesaurus* in Brazilian Portuguese, and single out the one we had in mind when we embarked on the compilation of our computerized lexical resourc. This specification turned out to be necessary for different specialists have used the term *thesaurus* to denote at least six different objets (Flexner, 1997; Lutz, 1994, Neufeldt, 1997; Roget, 1953):

1. An inventory of the vocabulary items in use in a particular language;
2. A thematically based dictionary, i.e., an onomasiologic dictionary;
3. A dictionary containing a store of synonyms and antonyms;
4. An index to information stored in a computer, consisting of a comprehensive list of subjects concerning which information may be retrieved by using the proper key terms;
5. A file containing a store of synonyms that are displayed to the user during the automatic proofreading process;
6. A dictionary of synonyms and antonyms stored in memory for use in word processing.

The Wordnet.Br core database is an instance of Object 6.

[1] Future work will include the specification of glosses for each *synset* and of hyponymy and meronymy relations between those synonym sets

2.2 Synonymy and Similarity of Meaning

The Wordnet.Br core database compilation process benefited from two key WordNet notions: the notions of synset and of lexical matrix. It is common ground that absolute synonyms are rare in language, if they exist at all. Thus, the notion of synset is derived from the conception of the symmetrical relation of meaning similarity, for "theories of lexical semantics do not depend on truth-functional conceptions of synonymy: semantic similarity is sufficient", and synonymy proper is understood as "simply one end of a continuum on which similarity of meaning can be graded" (Miller and Fellbaum, 1991, p.202).

2.3 Reusability of Published Dictionaries and the Reference Corpus

It is a fact that the compilation of a bulky dictionary is a time consuming activity and requires a team of more than fifty lexicographers, each responsible for (i) selecting the headwords which will head the dictionary entries, (ii) defining the number of senses for each headword, and (iii) exemplifying the senses with sentences and expressions from a selected corpus.

As a matter of fact,

- Dictionary entries specify a cluster of information: orthographical, phonological, etymological, morphological, syntactic, definitional, collocational, variational, register information about words, and sense relations such as synonymy and antonymy.
- Dictionaries extensively use the synonymy and antonymy word forms in their defining procedure to define headwords.

It is also a fact that lexicographers are aware that compiling dictionary entries involves making a very hard decision as to dealing with polysemy and homonymy. In other words, they have to decide on whether to lump or split word senses, or on whether to create fresh new entries for the same word form. Such decisions, however, are arbitrary, for lexicographers take their own personal experience and expertise to make their decisions; and probably that is the only way they manage to compile their unique store of words. Thus, reusing lexicographical information requires caution.

It must be stressed though that if we want to use dictionary lexicographical information in natural language processing projects, it must be mined and filtered carefully.[2]

The advent of computers have allowed lexicographers to use machine-readable, large-scale corpuses in their work, establishing procedures as follows (Stubbs, 2001): (a) to gather concordances from the corpus; (b) to cluster the concordances around nuclear sense clusters; (c) to lump or split nuclear clusters; (d) to encode the relevant lexical information by means of the highly-constrained language of dictionary definitions.

Given our small team of researchers, and the two-year time stipulated for the project, we bypassed those procedures and decided to reuse five outstanding published dictionaries of Brazilian Portuguese, which were chosen for the following reasons: (i) their being "fruits of the cumulative wisdom of generations of lexicographers", and their "sheer breadth of coverage" (just to borrow Kilgarriff's words, 1993, p.365); (ii) the relevant sense relations one of the five dictionaries registers can be complemented by similar pieces of information found in the other four; (iii) instead of using the Aristotelian analytical definition (i.e., *genus and differentiae*) to define word senses, they extensively use the synonymy and antonymy word forms in their defining procedure, feature that helped speed up the process of collecting and selelcting thousasnds of synonym and antonym word forms.

Two of them, Ferreira (1999) and Weiszflog (1998) are the most traditional and bulkier Brazilian Portuguese dictionaries. Their electronic versions speeded up further the process of synonym and antonym mining. Barbosa (1999) and Fernandes (1997) are specific dictionaries of synonyms and antonyms, and were used as complementary material. The fifth dictionary is a dictionary of verbs (Borba, 1990) that uses a Chafe-based semantic classification of verbs (Chafe, 1970). For each verb entry, the Borba's dictionary registers the relevant categories ("state", "action", "process", and "action-process"), its sense definitions, when available, its synonyms, its grammatical features, its potential argument structures, its selectional restrictions, and sample sentences extracted from corpora. Such specificity help fine tune the process of compiling the verb synsets.

---

[2] Acquiring such information is a hard problem and has been usually approached by reusing, merging, and tuning existing lexical material. This initiative has been frequently reported in the literature (see Kilgarriff, 1993, 1997, and the papers cited therein).

2.4 Dictionary Sense Distinctions and Leading Strategies

In the heart of the task of compiling dictionaries for the general public is the specification of word sense distinctions. On analyzing the LDOCE entries (Summers, 1995), Kilgarriff (1993, p.372-374) categorized four general types of sense distinctions made by lexicographers.

- "Generalizing Metaphors", i.e., a sense that is the generalization of a specific sense. For example:

  martelar (to hammer)
  sense 1: hit with a hammer (Core meaning)
  sense 2: insist (Generalizing meaning)

- "Must-be-theres", i.e., one of the senses is a logical consequence of the other. For example:

  casamento (marriage)
  sense 1: the event of getting married (Event)
  sense 2: the subsequent state of being married (Resulting state)

- "Domain Shift", i.e., a sense that extends the "original" sense to different domains. For example:

  leve (light)
  sense 1: not heavy, with little weight (Mass dimension)
  sense 2: nimble, agile" (Kinetic dimension)

- "Natural and social kinds", i.e., the different word senses apply to world entities or situations that have many attributes in common, but belong to different classes of things. For example:

  asa (wing)
  sense 1: a bird´s wing (Natural)
  sense 2: an airplane wing (Social)

Besides being aware of these sense distinctions, the following leading strategies were observed by our team of linguists:

- Checking whether particular grammatical or semantic features were necessary to lump together or to split over synonym sets (necessity strategy);

- Checking the symmetry property of both synonymy and antonymy (consistency strategy);
- Checking how wide the sense variation of a lexical unit were so that new senses would be posited (centrality strategy).

3. The Representation Domain

3.1 The Synset and the Lexical Matrix Constructs

The Wordnet.Br core database compilation process benefited from the two key WordNet constructs: the *synset* and the *lexical matrix*. It is common ground that absolute synonyms are rare in language, if they exist at all. Thus, the notion of synset is derived from the aforementioned conception of the symmetrical relation of meaning similarity.

Miller and Fellbaum (1991) argue that each synset is a set made up of semantically similar words that serve as unambiguous designators of meanings; they also assume that a speaker of a language has mastered collections of concepts and are expected to recognize them from the words that make up the synsets. The notion of lexical matrix, in turn, is intended to capture the "many to many" associations between form and meaning. In other words, it is conceived of as a mapping between written words, form representations, and synsets, meaning representations.

After adopting the key WordNet notions, the linguits embarked on the processes of mining synsets. The best way to understand how the compilers "mined" for synonyms into the reference corpus is to follow a real example.

Let us take, as our starting point of the mining process, the verb *lembrar* (English: "to remember"). Weiszflog (1998) distinguishes seven senses. After collecting the synonyms, and disregarding their definitions, the following synonym sets could be compiled:

1. {*lembrar*, *recordar*}
   (English: {"to remember", "to recall"})

2. {*lembrar*, *advertir*, *notar*}
   (English: {"to remember", "to warn", "to notify"})

3. {*lembrar*, *sugerir*}
(English: {"to suggest", "to evoke", "to hint"})

4. {*lembrar*, *recomendar*}
(English: {"to remember", "to commend"})

After that preliminary analysis, the linguist checked the consistency of the four synonym sets by looking up the dictionary synonym entries for the remaining five verbs: *recordar*, *advertir*, *notar*, *sugerir*, and *recomendar*.

Accordingly, the linguist, for example, looked up the dictionary entry for the verb *recordar*. Its first sense is given by the paraphrase *trazer à memória* (English; "to call back to memory"), and its fourth sense by the synonym *lembrar*. As these two senses are very close, and the examples confirm the similarity between the two, the synonym set 1 said to be consistent.

The very same process was repeated to every verb listed above until the list was exhausted. The analytical cycle began again by collecting the synonyms from the next dictionary entry in the alphabetical order.

It should be pointed up that, when the linguist analyzed the verb *esquecer* (English: "to forget"), the canonical Brazilian Portuguese antonym for *lembrar*, he found only one synonym for it: the verb *olvidar* (Vulgar Latin: "oblitare"; English: "to efface"). So, after the consistency analysis, the following synonym set was compiled:

5. {esquecer, olvidar}

The dictionary also registers this antonymy indirectly: *lembrar* and *esquecer* are defined by means of the paraphrases *trazer à memória* and *perder a memória de* (English: "to stop remembering"), respectively. Thus, the information was checked through cross-reference of entries and confirmed the antonymic pair (*lembrar*, *esquecer*), which stresses the importance of examining paraphrases carefully.

Just for the record: the synonym set (6) and its antonym set (7) are transcribed bellow:

6. {amentar2, comemorar, ementar, escordar1, lembrar, memorar, reconstituir, recordar, relembrar, rememorar, rever1, revisitar, reviver, revivescer, ver}
7. {deslembrar, desmemoriar, esquecer, olvidar}

3.2 The Wordnet.Br Core Database Design

Each Wordnet.Br core database entry consists of the following template:

[<Headword> n (<X>)

Sense n.1 [{Synset}; {Antonym Synset}]
...
Sense n.m [{Synset}; {Antonym Synset}]]

where n is the entry identification number; X is a noun, verb, adjective, or adverb; and n.1 ... n.m are sense identification numbers of the entry n.

From the logical point of view, the Wordnet.Br core database overall structure is made up of two lists: an Entry List (EL), the Wordnet.Br core database entries ordered alphabetically, and the Synset List (SL), the list of the synsets. Each element of a synset is necessarily an element of the EL. Each EL entry, besides being specified for its graphemic representation, it is also specified for a particular Sense Specification (SS). Each SS is indexed by three memory pointers: the "synonymy pointer" points to a particular synset (say synset 1) in the SL; the "antonymy pointer" points to a particular synset in the SL (when there is one) which is the antonym of the synset 1; and the "sense pointer", besides identifying the sense, say sense 1, points to the particular entry in the EL to which both synsets are associated.

Each synset in the SL is represented as "double-faced" list. One side lists specific elements of the EL that are members of the synset and the other side specifies a list SSs to which the synset is part. In other words, let us name the faces: the Entry Face (EF) and the Sense Face (SF). The EF contains pointers to the elements of the EL that are related to one another by means of the synonymy relation. The SF contains a list of SSs that indicates all SSs to which the synset is linked.

A conventional relational database management system is used. Its main functionalities include: the storage of general information of the Wordnet.Br core database and its bookkeeping. The design includes the complete loading of all entries and their related information. The key feature is the automatic entry generation. Once the synset is entered in the Wordnet.Br core database

or updated, the system generates the appropriate entries automatically. Just to illustrate with numbers: 3,872 verb synsets generate 10,204 verb entries.

4. The Computational Domain

4.1 The Editing Tool

The editing tool is a Windows®-based interface where the linguists enter synsets, sample-sentences, glosses, and generates diffrent lists (synsets listed by syntactic category, number of elements, degree of homonymy and polysemy, and list of sample sentences) and statistics (number of headwords, synsets, antonymous synsets, and headword/synset ratio).

4.2 The Wordnet.Br Core Statistics

CATEGORY	LEXICAL UNITS	SYNSETS
Verbs	~ 11,000	~ 4.000
Nouns	~ 17,000	~ 8.000
Adjectives	~ 15,000	~ 6,000
Adverbs	~ 1,000	~ 500
TOTAL	**~ 44,000**	**~ 18,500**

As the paper alerted care must be taken not to carry over published dictionary flaws into the Wordnet.Br core. But, despite their imperfections, the dictionaries selected as the reference corpus proved to be valuable resources of lexical-semantic information. Thanks to them, and to the systematic mining process and filtering strategies, the Wordnet.Br core database, with circa 18,000 synsets, can be further refined and updated to the Wordnet.Br. Accordingly, future steps will involve the specification of glosses for each sense, of sample sentences and expressions for each word form, and of the logical-conceptual relations of meronymy-holonymy and hyponymy-hypernymy.

References

Azevedo, F.F.S. (1983) *Dicionário Analógico da Língua Portuguesa*. Brasília: Thesaurus.

Barbosa, O. (1999) *Grande Dicionário de Sinônimos e Antônimos*. Ediouro, Rio de Janeiro.

Borba, F.S. (coord.) (1990) *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil*. Editora da Unesp, São Paulo.

Chafe, W. (1970) *Meaning and the Structure of Language*. The University of Chicago Press, Chicago.

Cruse, D.A. (1986) *Lexical Semantics*. Cambridge University Press, New York.

Dias-da-Silva, B.C. (1996). The technological facet of language studies: natural language processing, PhD Diss, FCL-UNESP, Araraquara, Brasil. (In Portuguese)

Dias-da-silva, B.C. (1998) Bridging the gap between linguistic theory and natural language processing. In: Caron, B. (ed.) *16th International Congress of Linguists*. Pergamon-Elsevier Science, Oxford 10 p.

Dias-da-Silva, B.C., Oliveira, M.F., Hasegawa, R., Moraes, H.R., Amorim, D., Paschoalino, C. Nascimento, A.C. (2000) A construção de um thesaurus eletrônico para o português do Brasil. In: *Proceedings of the 5th PROPOR - Encontro para o processamento computacional da língua portuguesa escrita e falada*, Atibaia, Brazil, p.01-10.

Dias-da-Silva, B.C., Oliveira, M. F., Moraes, H.R (2002) Groundwork for the Development of the Brazilian Portuguese Wordnet. In: Ranchhold, E.M.; Mamede, N.J. (eds.) *Advances in natural language processing*. Springer-Verlag, Berlin, p.189-196.

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

Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, Mass.

Fernandes, F. (1997) *Dicionário de Sinônimos e Antônimos da Língua Portuguesa*. Globo, São Paulo.

Ferreira, A.B.H. (1999) *Dicionário Aurélio Eletrônico Século XXI* (versão 3.0). Lexicon, São Paulo.

Flexner, S.B. (ed.) (1997): *Random House Webster´s Unabridged Electronic Dictionary* (Version 2.0).New York.,: Hndom House Inc.

Hayes-Roth, F. (1990). Expert Systems. In: E. Shapiro (ed.) *Encyclopedia of artificial intelligence*. New York: Wiley, p.287-98.

Kilgarriff, A. (1993) *Dictionary word sense distinctions: an inquiry into their nature*. Computer and the Humanities, 26, p.365-387

Kilgarriff, A. (1997) I don´t Believe in Word Senses. *Computer and the Humanities*, 31, p.91-113.
Kilgarriff, A., Yallop, (2000) C. "What's in a Thesaurus?". In: *Proceedings of the 2nd Conference on Language Resources and Evaluation*, Athens, Greece, 8 p.
Lutz, W.D (1994) *The Cambridge Thesaurus of American English*. Cambridge: Cambridge University Press.
Miller, C., Fellbaum, C. (1998) *Semantic networks of English*. *Cognition*, 41, p.197-229.
Nascentes, A. (1981) *Dicionário de Sinônimos*. São Paulo: Nova Fronteira.
Neufeldt, V. (ed.) (1997) *Webster´s New World Dictionary & Thesaurus* (Version 1.0). New York: Macmillan.
Roget, P.M. (1953) *Thesaurus*. Middlessex: Penguin Books. (Original ed. 1852)
Stubbs, M. (2001) *Words and phrases.* Oxford: Blackwell.
Summers, D. (ed.) (1995) *Longman Dictionary of Contemporary English*. Longman, Essex.
Weiszflog, W. (ed) (1998) *Michaelis português – moderno dicionário da língua portuguesa* (versão 1.1). DTS Software Brasil Ltda, São Paulo.

Human language technologies. Researching software and systems that bridge the linguistic divide between people and machines to make communicating with computers as natural as speaking with family and friends. Highlights. Layer Trajectory BLSTM: New evolution enhances speech recognition technology.Â This form contains a series of checkboxes that, when selected, will update the search results and the form fields. Currently selected items are under the "current selections" heading. Refine Results search results. The Portuguese Language Orthographic Agreement of 1990 (Portuguese: Acordo OrtogrÃ¡fico da LÃngua Portuguesa de 1990; European pronunciation:[ÉËˆkoÉ¾Ã°w É”É¾tuËˆÉ£É¾afiku Ã°É ËˆÉ«Ä©É¡wÉ puÉ¾tuËˆÉ£ezÉ], Brazilian pronunciation: [aËˆkoÉ¹.dÊŠ oÉ¹.toËˆÉ¡É¾a.fi.kÊŠËˆ dÉËˆ lÃ©.É¡wÉ poÉ¹.tuËˆÉ¡e.zÉ]) is an international treaty whose purpose is to create a unified orthography for the Portuguese language, to be used by all the countries that have Portuguese as their official language. It was signed in Lisbon, on 16 December 1990, at the end 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.)