WordNet.PT$_{global}$ – Extending WordNet.PT to Portuguese varieties

Palmira Marrafa[1], Raquel Amaro[2] and Sara Mendes[2]
Group for the Computation of Lexical and Grammatical Knowledge,
Center of Linguistics of the University of Lisbon
Avenida Professor Gama Pinto, 2
1649-003 Lisboa, Portugal
[1]palmira.marrafa@netcabo.pt [2]{ramaro,sara.mendes}@clul.ul.pt

Abstract

This paper reports the results of the WordNet.PT$_{global}$ project, an extension of WordNet.PT to all Portuguese varieties. Profiting from a theoretical model of high level explanatory adequacy and from a convenient and flexible development tool, WordNet.PT$_{global}$ achieves a rich and multi-purpose lexical resource, suitable for contrastive studies and for a vast range of language-based applications covering all Portuguese varieties.

1 Introduction

WordNet.PT is being built since July 1999, at the Center of Linguistics of the University of Lisbon as a project developed by the *Group for the Computation of Lexical and Grammatical Knowledge* (CLG).

WordNet.PT is being developed within the general approach of EuroWordNet (Vossen 1998, 1999). Therefore, like each wordnet in EWN, WordNet.PT has a general conceptual architecture structured along the lines of the Princeton WordNet (Miller *et al.* 1990; Fellbaum 1998).

For early strategic reasons concerning applications, this project is being carried out on the basis of manual work, assuring the accuracy and reliability of its results.

Aiming at using the Portuguese WordNet in language learning applications, among others, the starting point for the specification of a fragment of the Portuguese lexicon, in the first phase of the project (1999-2003), consisted in the selection of a set of semantic domains covering concepts with high productivity in daily life communication. The encoding of language-internal relations followed a mixed top-down/bottom-up strategy for the extension of small local nets (Marrafa 2002). Such work firstly focused on nouns, but has since then been extended to all the main POS, a work which has resulted both in refining information specifications and increasing WordNet.PT coverage (Amaro *et al.* 2006; Marrafa *et al.* 2006; Amaro 2009; Mendes 2009).

Relational lexica, and wordnets in particular, play a leading role in machine lexical knowledge representation. Hence, providing Portuguese with such a rich linguistic resource, and particularly Portuguese varieties not often considered in lexical resources, is crucial, not only to researchers working in contrastive studies or with the so-called non-standard varieties, but also to the general public, as the database is made available for consultation in the WWW through an intuitive and perspicuous web interface. Such work is also particularly relevant as the resulting database can be extensively used in a vast range of language-based applications, able to cover, this way, all Portuguese varieties.

This paper depicts the work developed and the results achieved under the scope of the WordNet.PT$_{global}$ project, funded by Instituto Camões, which, as mentioned above, aims at extending WordNet.PT to Portuguese varieties.

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 70–74,
Edinburgh, Scotland, UK, July 27–31, 2011.

2 The Data

Portuguese is spoken in all five continents by over 250 million speakers, according to recent studies, and is the official language of 8 countries: Angola, Brazil, Cape Verde, East Timor, Guinea Bissau, Mozambique, Portugal, and Sao Tome e Principe. Being spoken in geographically distant regions and by very different communities, both in terms of size and culture, Portuguese is naturally expected to show variation. Despite this, regional varieties are far from being equally provided with linguistic resources representing their specificities, as most research work is focused either on the Brazilian or the European varieties[1].

In the work depicted here we aim at contributing to reverse this situation, considering that this kind of resource is particularly adequate to achieve this goal, since it allows for representing lexical variation in a very straightforward way: concepts are the basic unit in wordnets, defined by a set of lexical conceptual relations with other concepts, and represented by the set of lexical expressions (tendentially all) that denote them. We have to antecipate the possibility of different varieties showing distinct lexicalization patterns, particularly some lexical gaps or lexicalizations of more specific concepts. This is straightforwardly dealt with in the WordNet model: once a system of relevant tags has been implemented in the database in order to identify lexical expressions with regard to their corresponding varieties, lexical gaps are simply encoded by not associating the tag of the variety at stake to any of the variants in the synset; specific lexicalizations, on the other hand, are added to the network as a new node and associated to the variety tag at stake.

Our approach consisted in extracting 10 000 concepts from WordNet.PT, and associating them with the lexical expressions that denote them in each Portuguese variety considered in this project. In order to accomplish this, we consulted native speakers from each of these varieties, resident in their original communities, and asked them to pinpoint the expressions used for denoting the aforementioned 10 000 concepts. Informants were selected by Instituto Camões among undergrad students in Portuguese studies and supervised by Portuguese lectors in each local university. Besides the European Portuguese variety, which is already encoded in WordNet.PT, specifications for six other Portuguese varieties were integrated in the database [2] : Angolan Portuguese, Brazilian Portuguese, Cape Verdean Portuguese, East Timorese Portuguese, Mozambican Portuguese and Sao Tome e Principe Portuguese. For each concept, several lexicalizations were identified and both the marked and unmarked expressions regarding usage information[3] were considered and identified.

2.1 Data selection

As mentioned above, our approach for enriching WordNet.PT with lexicalizations from all Portuguese varieties consisted in extracting 10 000 concepts from WordNet.PT and associating them to the lexical expressions which denote them in each variety.

domain	nouns	verbs	adjectives	proper nouns	total
art	422	14	83	0	519
clothes	467	62	74	0	603
communication	314	151	106	82	653
education	536	37	30	82	685
food	1131	130	115	0	1376
geography	281	0	166	200	647
health	1159	92	175	0	1426
housing	595	28	46	0	669
human activities	641	0	0	0	641
human relations	620	189	100	0	909
living things	1597	113	119	1	1830
sports	480	34	23	2	539
transportation	659	562	67	30	659
all domains	7893	802	1022	284	10001
domain overlap	10,36%	12,54%	4,22%	22,62%	10,35%

Table 1: concepts extended to Portuguese varieties

[1] Official standardized versions of East Timorese and African varieties of Portuguese essentially correspond to that of European Portuguese. Moreover, these varieties are not provided with dedicated lexical resources, such as dictionaries or large-scale corpora. Being so, speakers in these regions generally use European Portuguese lexical resources, which only exceptionally cover lexical variants specific to these varieties.

[2] All Portuguese varieties spoken in countries where Portuguese is the official language were considered. However, for the time being, data from Guinean Portuguese are not yet encoded in the WordNet.PT$_{global}$ database due to difficulties in maintaining a regular contact with the native speakers consulted. Despite this, we still hope to be able to include this variety in the database at some point in the future.

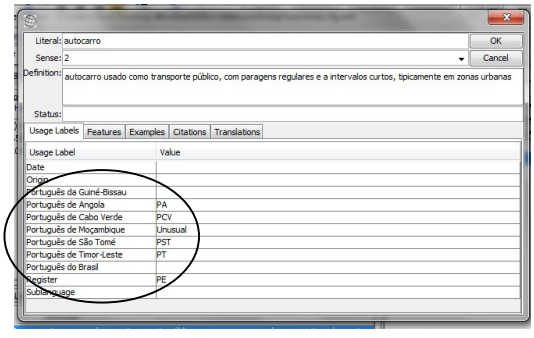
[3] Informants were provided with a limited inventory of usage markers: slang; vulgar; informal; humerous; popular; unusual; regional; technical; old-fashioned.

The semantic domain approach initially used in developing WordNet.PT, provided us with a natural starting point for the selection of data to be considered in this project. The table above presents the distribution, per POS and semantic domain[4], of the WordNet.PT concepts extended to non-European Portuguese varieties.

2.2 Data implementation

Once the data described above were presented to the native speakers consulted and their input organized, all the information obtained was incorporated in the database.

This way, for a concept like ***bus*** (public transportation which has regular pre-established stops at short intervals, typically operating within cities), for instance, the following lexicalizations were obtained: *autocarro*, *machibombo*, *machimbombo*, *ônibus*, and *microônibus*. *Autocarro* was found to be the more common expression used for denoting the concept at stake in Angola, Cape Verde, East Timor, Portugal, Sao Tome e Principe and Mozambique. However, this variant is marked as "unusual" in Mozambique variety. *Machibombo* and *machimbombo* are only used in Mozambique, whereas *ônibus* and *microônibus* are only used in Brazil. With this kind of data at hand, each lexicalization was tagged with regard to the varieties in which it is used and, for each variety, associated, when relevant, to a usage label, as illustrated below.



In the codification of the aforementioned information we used ***Synsetter*** – a new, very flexible wordnet development tool previously developed for the full implementation of innovative research results in WordNet.PT. In order to do so, this computational tool has been developed to straightforwardly allow for updates and improvements. In the specific case of the task addressed in this paper, extending the coverage of the WordNet.PT database to lexicalizations of different Portuguese varieties involved the design of additional features regarding the identification of Portuguese varieties and variety-dependent usage label encoding.

2.3 The results

Encoding the data obtained in WordNet.PT$_{global}$ extends a relevant fragment of WordNet.PT to Portuguese varieties other than European Portuguese. This way, researchers are provided with a crucial database for developing contrastive studies on the lexicon of different Portuguese varieties or research on a specific Portuguese variety, just to mention a possible application. The table below presents the distribution of variants per variety in the fragment of the lexicon considered, making apparent, for instance, that in the collection of data considered in this project some varieties have more synonym forms for denoting the same concept than others (see average of variants per concept).

Portuguese varieties	**number of concepts**	**number of variants**	**variants per concept (average)**
Angola	10 000	11713	1,17
Brazil	10 000	12060	1,20
Cape Verde	10 000	12563	1,26
East Timor	10 000	12131	1,21
Mozambique	10 000	11740	1,17
Portugal	10 000	13006	1,30
Sao Tome e Principe	6981	9552	1,37
all varieties	10 000	14751	1,47

Naturally, this is only an overall view of the results obtained. The new extended WordNet.PT version is also a crucial resource allowing for contrastive studies on lexicalization patterns depending on semantic domains or on frequency of use, for instance, for all or for specific Portuguese varieties.

In order to make these data publicly available, a new WordNet.PT version, the WordNet.PT$_{global}$ has been released on the WWW [5]. Releasing the WordNet.PT fragment extended to Portuguese

[4] Note that some of the concepts considered are associated to more than one semantic domain. This results in partial overlaps between semantic domains, whose extent is presented in the last row of Table 1.

[5] http://www.clul.ul.pt/wnglobal.

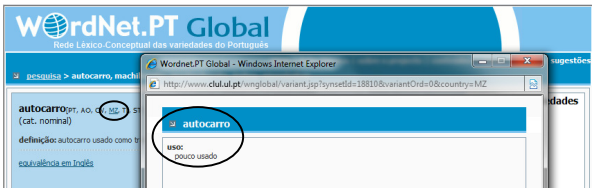
varieties online involved developing an updated version of the web interface for wordnet online navigation. In Section 3 we present the main features of this web interface and how users can navigate and straightforwardly access the data on Portuguese varieties.

3 Navigating the lexicon of Portuguese varieties online

The new updated version of the web interface for wordnet navigation was developed with specific features allowing for the visualization of information on Portuguese varieties and for narrowing down searches depending on the needs of the user. Among the most salient aspects of the new web interface we underline the following: allowing the user to restrict the search to a given (or to a set of) Portuguese variety(ies) (see caption below); displaying information about each lexical expression regarding the varieties which use it and whether this use is marked or not.

Going back to the example mentioned in section 2.2, in Portuguese, the concept {**bus**} can be denoted by several expressions, depending on the variety considered. This information is straightforwardly displayed and made available to the user by a simple system of tags, as illustrated below.

Also, all marked uses are indicated by underlining the variety label corresponding to the variety in which the use of the relevant expression is marked (see tags associated to *autocarro* in the caption above, particularly the *MZ* tag signaled by an arrow). By clicking on this label the relevant usage label is displayed, as illustrated below.



4 Final Remarks

WordNet.PT$_{global}$ is, thus, a relational lexicon allowing for modelling the lexicon of Portuguese varieties in a consistent and motivated way.

Covering 10 000 concepts, lexicalized by a total of 14 751 expressions representing all the main POS (nouns, verbs, adjectives, and proper nouns), WordNet.PT$_{global}$ also provides a lexical-conceptual network of relations establishing the relevant links between each concept and the other concepts in the net, in a total of more than 30 000 relations, including relations with their corresponding lexicalizations in English.

This way, Portuguese now has a rich and useful lexical resource covering all of its varieties (Angolan, Brazilian, Cape Verdean, East Timorese, European, Mozambican, Sao Tome e Principe and Guinean Portuguese (forthcoming – see footnote 1)), freely available for online consultation both to researchers and to the general public.

Moreover, the database presented in this paper can be extensively used in a vast range of language-based applications which are now able to cover all Portuguese varieties. As a final remark on future work, the data resulting from WordNet.PT$_{global}$ can be used as a basis for comparative studies regarding, for instance, variant distribution per variety. Note, however, that pursuing such studies requires comparable corpora for each variety, both with POS tagging and semantic annotation. Nonetheless, several advances are being taken in this direction[6].

[6] see http://www.clul.ul.pt/en/research-teams/87-linguistic-resources-for-the-study-of-the-african-varieties-of-portuguese-r.

References

Amaro, R. (2009) *Computation of Verbal Predicates in Portuguese: relational network, lexical-conceptual structure and context – the case of verbs of movement*, PhD dissertation, University of Lisbon.

Amaro, R., R. P. Chaves, P. Marrafa & S. Mendes (2006) "Enriching WordNets with new relations and with event and argument structures", *Proceedings of CICLing 2006*, Mexico City, pp. 28-40.

Fellbaum, C. (1998) (ed.) *WordNet: an Electronic Lexical Database*, MA: The MIT Press.

Marrafa, P. (2002) "The Portuguese WordNet: General Architecture and Semantic Internal Relations", *DELTA*, Brasil.

Marrafa, P., R. Amaro, R. P. Chaves, S. Lourosa & S. Mendes (2006) "WordNet.PT new directions", *Proceedings of GWC'06*, Jeju Island, pp. 319-321.

Mendes, S. (2009) S*yntax and Semantics of Adjectives in Portuguese: analysis and modelling*, PhD dissertation, University of Lisbon.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross & K. J. Miller (1990) "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, volume 3, number 4.

Vossen, P. (1998) (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Kluwer Academic Publishers.

Vossen, P. (1999) *EuroWordNet General Document*, University of Amsterdam.