PORTUGUESE WORDNET: GENERAL ARCHITECTURE AND INTERNAL SEMANTIC RELATIONS

(WordNet do Português: Arquitetura Geral e
Relações Semânticas Internas)

Palmira MARRAFA
(Universidade de Lisboa. Faculdade de Letras, Departamento de Linguística Geral e Românica and CLG-Computation of Lexical and Grammatical Knowledge Research Group – CLUL)

ABSTRACT: *In this paper we describe the general architecture and the main structuring relations of the Portuguese WordNet, under construction within the WordNet.PT project [*]. We also present empirical and technical motivations for information encoding, focusing on generic criteria concerning the expression of polysemy. For an overview of the first results of this project we present some statistical data, which show the representativeness of the major categories and relations.*
KEY-WORDS: *Synset; Lexical-Semantic Relations; Semantic Network; Natural Language Processing.*

RESUMO: *Neste artigo descreve-se a arquitetura geral da WordNet do Português em desenvolvimento no quadro do projecto WordNet.PT [*], bem como as relações que definem a sua estrutura básica. Aduz-se igualmente motivação empírica e técnica relativa à codificação da informação, com particular incidência nos critérios subjacentes à expressão da polissemia. É ainda apresentada uma estatística que proporciona uma perspectiva global dos primeiros resultados do projeto e que evidencia a representatividade das principais categorias e relações que constam da rede.*
PALAVRAS-CHAVE: *Synset; Relações Léxico-Semânticas; Rede Semântica; Processamento de Linguagem Natural.*

[*] The WordNet.PT project is funded by Instituto Camões.

0. Introduction

The work presented here concerns the first results of the ongoing WordNet.PT project*, which aims at constructing a large-scale wordnet for Portuguese. This project is being developed in the EuroWordNet framework.

EuroWordNet is a multilingual database that includes wordnets for several European languages, mainly structured along the same lines of the Princeton WordNet and interconnected via an Inter-Lingual-Index. The overall design of this model is described in section 2.

Language-specific wordnets are built as relatively independent systems. That makes it possible to adopt different tools and methodologies. WordNet.PT options on this matter are presented in section 3.

Despite the specific properties of individual wordnets, their compatibility is ensured by the assumption of the same interpretation of lexical-semantic internal relations that form their structuring backbone. In order to achieve that goal all the relations are defined on the basis of explicit tests. Yet, encoding specific semantic relational information is not a trivial task. Section 4 discusses the main issues concerning the major relations used in the Portuguese WordNet at the present phase.

The EuroWordNet database allows adding features to the defined relations to precise their semantic implications. The features used in the Portuguese WordNet are illustrated in section 4 as well.

Section 5 presents a global quantitative overview of the WordNet.PT project results at the current stage.

In the last section we briefly conclude by pointing out the scientific and technical consistency of WordNet.PT and its potential impact both from a theoretical and a language engineering point of view.

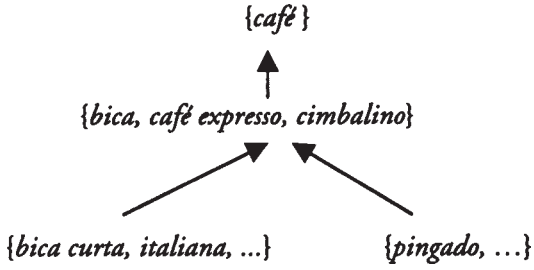
1. General model

WordNet.PT is being developed within the general framework of EuroWordNet (Vossen (1998 and 1999)).

EuroWordNet is a multilingual database consisting of individual wordnets for several European languages. The individual wordnets are basically structured along the same lines of the Princeton WordNet (Miller *et al*. (1990), Fellbaum (1998a and 1998b)).

A wordnet is a semantic network, in which the basic units are concepts, represented by sets of synonyms (synsets). Each synset contains the expressions (single words or complex sequences) that are lexicalisations of the same concept. For instance, the Portuguese expressions *bica, café expresso, cimbalino* are included in the same synset, since all of them are lexicalisations for the same concept (lexicalised in English by *espresso*).

Wordnets are structured on the basis of lexical and conceptual relations. The meaning of the lexical units is not defined by means of paraphrases, as usually in conventional dictionaries. It is rather derived from their relations with the other members of the same synset and from their lexical-semantic relations with other synsets, as illustrated below:



The meaning of *bica* is partially[1] derived from the synonymy relation with the other expressions inside the synset and from the conceptual relations with the synset {*café*}, which represents a more general concept, and with the synsets {*bica curta, italiana,..*} and {*pingado,..*}, which represent more specific concepts.

In a certain sense, meaning is constructed. It emerges from the structure of the network.

[1] The relations encoded for this item are not exhaustively represented here.

Complementary information is provided by glosses associated to synsets.

The monolingual wordnets are interconnected via the so-called Inter-Lingual-Index (ILI).

The ILI is an unstructured list of synsets (called ILI-records), each one representing a concept, in the same way as in individual wordnets. The ILI-records are mainly taken from WordNet 1.5.

Language-specific synsets linked to the same ILI-record should be cross-linguistically equivalent.

The most important equivalence relation is synonymy equivalence (eq_synonmy). It links a language-specific synset and an ILI-record that express the same concept (e.g. *bica* is connected by eq_synonmy to the ILI-record that includes *espresso*).

Whenever there is no ILI-record to which a given language-specific synset can be properly linked by means of the synonymy equivalence relation, complex equivalence relations can be specified, as for instance the equivalence to a more generic or to a more specific concept. In this situation, more than one equivalence relation can be specified for the same synset, if necessary. The option in the Portuguese WordNet is, as much as possible, for fine-grained specifications. This way, fuzziness can be reduced in the case of equivalence relations other than the very straightforward synonymy equivalence. Simultaneously, more reliable results can be achieved with regard to inter-lingual comparison[2].

It is worth noticing that the WordNet.PT ILI equivalence relations correspond mostly to synonymy relations (cf. section 5). This has great relevance for multilingual applications.

2. Resources, tools and methodology

The linguistic resources available for Portuguese being not suitable enough to the purpose of building a wordnet, the WordNet.PT project is being carried out mainly on the basis of manual work.

[2] See Peters *et al.* (1998) for further motivations of complex equivalence relations.

The building tool adopted is Polaris, the general EuroWordNet database tool.

As for other tools, the WordNet.PT team developed a browser/web interface, WebNet, and a search engine, Detonador[3], for quantifying the network and executing useful queries about the different types of relational structures within the database.

From a methodological point of view, two different main approaches are followed within EuroWordNet for encoding semantic relations, as pointed out by Vossen *et al*. (1998): the Merge model and the Expand model.

In the first case, the development of synsets and their internal semantic relations is independent of WordNet 1.5. Afterwards, the equivalence relations to WordNet 1.5 are generated. Such wordnets are independent of WordNet 1.5 and maintain language-specific properties.

In the Expand model the WordNet 1.5 synsets are translated into the other language and the WordNet 1.5 relations are taken over and adapted to EuroWordNet. Therefore, the resulting wordnets are very close to WordNet 1.5.

WordNet.PT uses a mixed approach:

- independent selection of vocabulary
- independent development of language-internal relations
- Princeton WordNet as a main source of inspiration

Since we aim at using the Portuguese WordNet in language learning/teaching applications, among others, the starting point for the specification of a fragment of the lexicon was the selection of a set of semantic domains covering concepts with high productivity in daily life communication.

Although the main source of inspiration for selecting the candidate nodes in each semantic domain is WordNet 1.5, the starting point for the definition of the core-wordnets in each domain is not its translation. This way, the Portuguese-specific properties can be captured.

[3] Denotador was developed by Rui Pedro Chaves.

The encoding of language-internal relations is manually done, following a mixed top-down/bottom-up strategy for the extension of small local nets.

Synsets include not only single words, but also non-atomic expressions (e.g. *café expresso*). The elementary units in wordnets are concepts, as already pointed out, and concepts can be lexicalised by one or more words.

Since the different parts-of-speech are not organized on the basis of the same lexical-semantic relations, building wordnet fragments per category is less time-consuming and increases accuracy in contexts of limited resources, at least in a first phase.

Accordingly, the first stage of WordNet.PT has mostly focused on nominal lexicalisations.

3. Internal Relations

3.1. Synonymy

As referred to before, synonymy is the most basic relation in wordnets: concepts are represented as synsets (sets of synonyms). As a consequence, the lexical and conceptual-semantic relations that structure the database link synsets to other synsets.

As well known, synonymy is not an uniform concept[4]. The strongest one presupposes absolute or true synonyms, which can be informally defined as follows:

> *Two expressions are synonyms if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made*.

Nevertheless, true synonyms are rare, if any. Accordingly, as usually in wordnets (Princeton WordNet and EuroWordNet), we use a weaker definition ("semantic similarity") as stated by Miller *et al.* (1990):

> "*two expressions are synonyms in a linguistic context C if the substitution of one for the other in C does not alter the truth value of C*".

[4] For a discussion of synonymy varieties and their characteristics see, for instance, Cruse (1986).

Symmetry is required:

if A is a synonym of B in C, then B is a synonym of A in C.

This requirement is subject to verification by means of an explicit test:

Test 1

A is a B/B is an A

Let us consider some examples:

(1) a. Uma *bica* é um *café expresso*. True
'A "bica" is an espresso'
b. Um *café expresso* é uma *bica*. True
'An espresso is a "bica"'

(2) a. Uma *bica* é um *café*. True
'A "bica" is a coffee'
b. Um *café* é uma *bica*. False
'A coffee is a "bica"'

Taking into account the definition above in conjunction with the symmetry constraint, *bica* and *café expresso* belong to the same synset, but *bica* and *café* do not.

Items that do not satisfy Test 1, above, but do satisfy Test 2, below, are not synonyms. They are near synonyms.

Test 2

A is a kind of B/ B is a kind of A

Near synonymy is also expressed, but as a distinct relation. In other terms, near synonyms do not belong to the same synset.

Polysemy. Regarding polysemy, it has not an uniform treatment in WordNet.PT, since, as well known, polysemy is far from being an uniform phenomenon.

Avoiding over- and under-differentiation of senses is the general goal on this matter. However, splitting or collapsing multiple senses has often weak motivation. Besides, the optimal degree of specification of meaning in computational lexica depends to a large extent on the goals of the applications.

We have adopted a balanced approach, mainly based on the well-known indicators of discreteness[5] presented below[6].

• **Identity effect**. A word form having senses with a high degree of discreteness does not support different readings in elliptical contexts, as illustrated in (3):

(3) Ele não tem carneiro e ela também não.

'He not has lamb and she also not'

"He does not have (a) lamb and neither does not she."

It cannot be the case that the overt occurrence of *carneiro* is interpreted as a kind of meat and the covert one as a kind of animal.

Independent truth conditions. Distinct senses of a word form induce independent truth conditions:

(4) Ele não tem carneiro.

'He not has lamb'

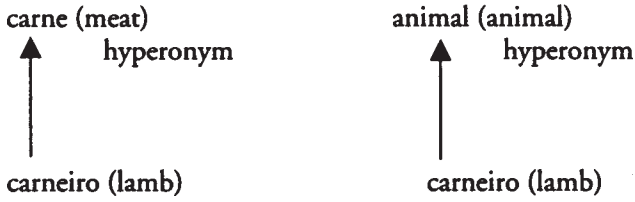
"He does not have (a) lamb."

[5] Along the lines of several previous studies (as for instance Cruse (2000)), we do not assume a polysemy/monosemy dichotomy. We rather presuppose a gradient of discreteness (for a different position see Geeraerts (1993), among others).

[6] These indicators are usually called "ambiguity tests". Nevertheless, as pointed out by Cruse (1995 and 2000), ambiguity requires antagonism and the indicators referred to above are not reliable for diagnosing full antagonistic readings. Anyway, the aim here is just determining a relevant degree of discreteness, not necessarily full antagonism.

The truth value of (4) varies in function of the different readings of *carneiro*, which is an indicator of discreteness.

• **Independent lexical-conceptual relations**. Senses of a word form having a high degree of discreteness do not belong to the same semantic local domain. They are related to different lexical-conceptual units, as exemplified:



• **Definitional distinctness**. There is no uniform definition of the meaning of a word form that encompasses senses with a high degree of discreteness and, at the same time, semantically distinguishes it from all the other meaning forms.

The criteria referred to above are conjointly considered.

Besides, productivity of alternations, both in Portuguese and cross-linguistically, is also taken into account. The lack of a systematic relation seems to be an additional indicator of discreteness.

Moreover, avoiding both over-differentiation and under-differentiation is still a hard task, since tests are not always conclusive, relations between senses splitted into different synsets are often very difficult to capture and the relevance of the degree of differentiation varies from application to application.

3.2. Hyponymy/Hyperonymy

The hyponymy/hyperonymy relation (also called subordination/superordination, subset/superset, ISA relation) is the most fundamental structuring relation in wordnets. It can be informally defined as follows:

α *is a hyponym of* β *(*β *hyperonym of* α*)*

iff

(i) α *is a kind of* β

and

(ii) β *is not a kind of* α

Let us consider some examples:

(5) a. Uma *bica* é um/um tipo de *café*. True

'A "bica" is a/a kind of coffee'

b. Um *café* é uma/um tipo de *bica*. False

'A "coffee" is a/a kind of 'bica'"

The data in (5) show that there is a hyponymy/hyperonymy relation between *bica* and *café*: *bica* is more specific than *café*.

Using just the test above does not prevent intermediate levels in the hierarchy from being skipped, as showed below:

(6) a. Uma *bica* é uma/um tipo de *bebida*. True

'A "bica" is a/a kind of drink'

b. Uma *bebida* é uma/um tipo de *bica*. False

The result of the test is the same for the pair *bica* and *bebida* and for the pair *bica* and *café.*

To avoid this situation, co-hyponymy is required to establish a genuine hyponymy relation (as stated by Vossen (1999)):

> *"If a pair of words W1 and W2 fits the test frame*
> *then there should be at least one other word W3*
> *which fits this frame in relation to W2 so that W1*
> *and W3 are so-called co-hyponyms of W2.*

The decision procedure involves linguistic tests, mostly depicted in Vossen (1999).

Complementarily, an informal qualia roles based "measure" is used. Hyponymy is determined upon the values specified for the qualia roles shared by hyponyms and hyperonyms.

Moreover, synsets are described by glosses, which have the format of an informal definition that includes the immediately dominating hyperonym plus the specific properties of each synset. This format enables incorrect specification of hyponymy to be detected.

To avoid non-disjoint co-hyponymy, or false co-hyponymy, non-lexicalised expressions can be encoded: (i) whenever a group of synsets (potentially co-hyponyms) lack lexicalised hyperonyms at all; (ii) whenever, otherwise, a group of synsets lacking true co-hyponymy would be represented as co-hyponyms, due to the fact that some of them have an intermediate non-lexicalised hyperonym. However, WordNet.PT contains mostly lexicalised expressions, non-lexicalised expressions being encoded just as a last resort.

3.3. Meronymy/Holonymy

The part/hole relation is another major relation coded in wordnets[7]. As pointed out in several studies (Cruse (1986), Winston *et al*. (1987), Vossen and Copestake (1993), Vossen (1995 and 1998), among others), the part/whole relations form a complex family of relations. Along the same lines as EuroWordNet, WordNet.PT distinguishes five part/whole relation subtypes:

- mero_part/holo_part (part/whole):
  dedo ("finger")/*mão* ("hand")
- mero_member/holo_member (member/set):
  jogador ("player")/*equipa* ("team")
- mero_location/holo_location (location/place):
  baixa ("down town")/*cidade* ("town")

[7] "The second major type of branching lexical hierarchy is the part-whole type" (Cruse (1986:157)).

- mero_portion/holo_portion (portion/whole):
  fatia ("slice")/*bolo* ("cake")
- mero_made_of/holo_made_of (substance/object):
  papel ("paper")/*jornal* ("newspaper")

As a consequence of the design of the database all the relations are interpreted as bi-directional relations. In other terms, for every relation stated in one direction a reversed counterpart is automatically generated.

Conceptual bi-directionality is mostly the canonical situation. However, in a few situations it does not hold. Meronymy/holonymy dependences are a typical case of variation, as illustrated below:

(7) *braço* ("arm") has_holonym *mão* ("hand")
mão ("hand") has_meronym *braço* ("arm")

(8)a. *carro*("car") has_meronym *porta* ("door")
porta ("door") has_holonym *carro* ("car")

Whereas in (7) the relation holds canonically in both directions, in (8)a. it does not. As a matter of fact, *carro* is not a canonical holonym of *porta*.

Adding features to the defined relations makes it possible to precise their nature and semantic implications. The differentiation between canonical and non canonical part/whole relation is specified by adding the feature *reversed* in the last case:

(8)b. *carro* ("car") has_meronym *porta* ("door")
porta ("door") has_holonym *carro* ("car") *reversed*

The feature *reversed* means that the tagged counterpart of the relation is automatically generated by the system. It is not an implication of the relation itself.

Whenever a given holonym has disjoint groups of meronyms, the features conjunction and disjunction are used to precise the relations.

The feature negation is used to express that the relation does not hold.

Adding features allows for introducing fine-grained relation distinctions, which are very important for using WordNet.PT in inference systems and Portuguese learning/teaching applications, among others.

3.4. Function relations

Besides the major relations depicted in the previous sections, the database also includes function relations, which cover several aspects of semantic entailment. They are mainly used to encode information on the participants (arguments or adjuncts) typically involved in a given event and strongly 'implied/incorporated' in the meaning of a lexical unit. They are generically referred to as involvement (and co_involvement) relations. Let us see some examples:

- involved_instrument
  telefonar ("to telephone") /*telefone* ("telephone")
- involved_agent
  jogar ("to play")/*jogador* ("player")
- co_agent_instrument
  guitarrista ("guitar player")/*guitarra* ("guitar")
- co_agent_result
  poeta ("poet")/*poema* ("poem")

Function relations have strong cognitive motivation: function seems to be a major feature of the organization of human knowledge[8].

The information encoded under this type of relations allows for capturing incorporation patterns in a language and incorporation patterns differences across languages. Hence, it is very useful both from the theoretical point of view and for applications building.

[8] On this matter, see, for instance, Vossen (1999).

4. Results (version 1. 0)

Due to operative and resource economy reasons, on the one hand, and to applications goals, on the other hand, the first version of the Portuguese WordNet is mostly focused on nouns. It contains approximately 10000 distinct noun forms within a total of about 11000 word forms.

Extending the coverage to verb and adjective forms is presently starting.

As referred to before (cf. 4. 2), nouns are organized prominently in terms of the hyponymy inclusion relation, meronymy being the other major relation. As a consequence, internal relations other than hyponymy and meronymy have irrelevant statistical values in this phase and are neglected in the table below, which provides a global quantitative overview of the WordNet.PT current stage, both in extension and depth terms.

Statistics	**Nouns**	**Verbs**	**Adjectives**
Synsets (Word-senses)	8100	424	291
Wordforms	9813	633	485
Average of Wordforms per Synset	1.21	1.49	1.67
Glossed Synsets	7235	332	219
Hyponymy Relations	8357	393	0
Meronymy Relations	2781	0	0
Sub_event Relations	0	53	0
ILI Relations	8089	729	260
ILI *Synonymy* Relations	5281	250	206

Table 1: Data Extracted from WN.PT 1.0

The WordNet.PT results will be publicly available at Centro de Linguística da Universidade de Lisboa and Instituto Camões websites.

5. Conclusion

WordNet.PT is a basic resource for a wide range of, both monolingual and multilingual, Computational Linguistics purposes and Language Engineering applications.

From a theoretical point of view, the Portuguese WordNet is being developed in a quite rich framework, allowing for: (i) a psychologically plausible organization of the lexicon; (ii) yielding language-specific patterns and cross-linguistically generalizations.

Dealing with practical issues as well, a continuous evaluation of the relations of the model to be used is being done, in order to guarantee the meaning modelling is sufficiently motivated, provides a fine-grained characterization of lexicalisation patterns and is useful for the intended applications.

A large-scale linguistic database such as WordNet.PT opens quite challenging possibilities for Portuguese within the domains of Natural Language Processing and Language Technologies.

Ackowledgements

The WordNet.PT project has been made possible by the funding of Instituto Camões, which I thank in the person of its President, Jorge Couto.

I also thank the CLG-Computation of Lexical and Grammatical Knowledge Research Group for being with me in this project. A special thank to Rui Pedro Chaves and Sara Mendes.

I want to express my gratitude to Christiane Fellbaum and Piek Vossen for helping and supporting me in this project, whenever I have needed. I also thank them for their work, the most important source of inspiration for building WordNet.PT.

I am indebted to Heronides Moura and Marco Rocha for having given me the opportunity to present and discuss the first results of WordNet.PT at the Conference "Polysemy and Semantic Indetermination", Florianópolis, Brazil.

REFERENCES

CRUSE, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

_____. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In: P. Saint-Dizier and E. Viegas (eds.). *Computational Lexical Semantics*. Cambridge: Cambridge University Press.

_____. 2000. Aspects of the Micro-structure of Word Meanings. In: Y. Ravin and C. Leacock (eds.) *Polysemy: Theoretical and Computational Approaches*. Oxford: Oxford University Press.

Fellbaum, C. 1998a. A Semantic Netwok of English: The Mother of All WordNets. In: P. Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Fellbaum, C. 1998b. A Semantic Netwok of English Verbs. In: C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. MA: The MIT Press.

Geeraerts, D. 1993. Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4.3.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross & K. J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3 (4): 235-244.

Peters, W., P. Vossen, P. Díez-Orzas & G. Adrians. 1998. Cross-linguistic Alignement of WordNets with an Inter-Lingual Index. In: P. Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Vossen, P. 1995. Conceptual and Grammatical Individuation in the Lexicon. PhD thesis. University of Amsterdam. *Studies in Language and Language Use*, 15.

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

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

_____. 1999. *EuroWordNet General Document*. University of Amsterdam.

Vossen, P. and A. Copestake. 1993. Untangling Definition Structure into Knowledge Representation. In: E. J. Briscoe, A. Copestake & V. de Paiva (eds.) *Default Inheritance in Unification-Based Approaches to the Lexicon*. Cambridge: Cambridge University Press.

Vossen, P. *et al.* 1998. *EuroWordNet Tools and Resources Report*. University of Amsterdam.

Winston M., R. Chaffin & D. Herrmann. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11.