Antonymy in Brazilian Portuguese Descriptive Adjectives: an Analysis Proposal

Cláudia Dias de Barros[1], Oto Araújo Vale[1]

[1] PPGL - Universidade Federal de São Carlos, Rodovia Washington Luís, km 235 – SP-310, 13.565-905, São Carlos, São Paulo, Brazil
claudias84@gmail.com, otovale@ufscar.br

Abstract: This paper aims at presenting a study about the semantic relation of antonymy in Brazilian Portuguese (BP) adjectives, in order to contribute to the refinement of *WordNet.Br* (WN.Br). The methodology is presented too, and the most important step is the establishment of the direct and indirect antonymy of the 100 most frequent adjectives extracted from a BP corpus. At the end, some results and conclusions are shown.

Keywords: Adjectives, Antonymy, Lexicon, Semantics, WordNet

1 Introduction

The development of *wordnets* in many languages is a comprehensive research field in Natural Language Processing (NLP). It started in 1985 with the *WordNet* of *Princeton University* (WN.Pr) [1]. It is a lexical database, and its design is inspired by current psycholinguistic theories of human lexical memory. English nouns, adjectives, verbs and adverbs are organized into synonym sets (*synsets)* related by conceptual relations like hyponymy, meronymy and antonymy, for example.

The project of a Brazilian Portuguese wordnet is being developed since 2002 and is called *Wordnet.Br* (WN.Br) [2]. Currently, WN.Br core database presents the figures in Table 1:

Table 1. The WN.Br Core Statistics (extracted from [2], p. 302)

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

A portion of the WN.Br database can be accessed through TeP - *Electronic Thesaurus to Brazilian Portuguese* [3]. It is a BP electronic dictionary of synonyms and antonyms.

The antonymy presentation at TeP (Fig. 1) is not similar to the one presented at WN.Pr (Fig. 2), mainly because it does not show the representation of the indirect antonymy (by virtue of the semantic similarity to adjectives that do have direct antonymy) like 'obeso=gordo/magro' (*obese=fat/thin),* as WN.Pr does.

Fig. 1. TeP web interface

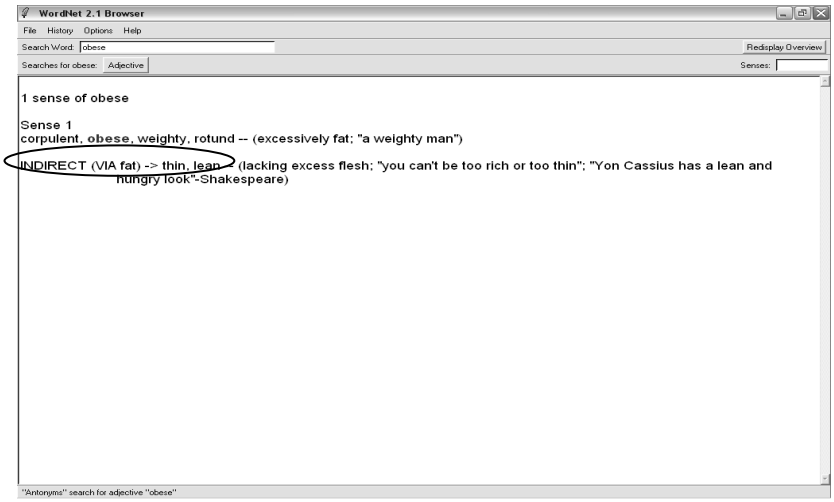

Fig. 2. WN.Pr offline version interface

This is the reason why this research aims at doing a refinement of the indirect antonymy representation at TeP, contributing to the WN.Br project's improvement and to the study of antonyms, for it is the most important relation among the adjectives.

1.1 Adjectives in WordNet

The main property of adjectives is to modify nouns, giving them a quality ('*beautiful* girl', for example).

Wn.Pr divides the adjectives into two major categories [1]: a) *descriptives*: they ascribe to their head nouns values of (typically) bipolar attributes and organize them in terms of antonymy, e.g. *big/little.* They combine with nouns to express some qualities of the thing, person or concept they designate; b) *relationals*: related to nouns, they can be changed in the expression '*preposition + noun'* and they don't have direct antonyms, e.g. '*musical* concert' - 'concert *of music'*.

There is another category of adjectives: *determinatives* – they occur just before the noun in Portuguese (and English), can be in other word classes like articles and pronouns, and they have no antonyms, e.g. '*certain* words'. WN.Pr doesn't present this adjectival category.

1.2 Antonymy

Antonymy is the basic semantic relation of descriptive adjectives [1]. According to [4], antonymy has some properties like, e. g., *markedness* (there is a marked and an unmarked term).

[5] divides the antonymy into four basic types: 1) *complementary* (ungradable opposites – *male/female);* 2) *contrary* (gradable[1] opposites – *hot/cold);* 3) *directional* (*up/down*); 4) *converse* (*sell/buy*).

WN.Pr divides the antonymy into two kinds:

1) *direct* or *lexical* – it is also called *canonic* antonymy [4], e.g. *black/white, good/bad.* It occurs between the word's forms and is basically formed by pairs of words that co-occur in several phrases, something that can be noticed in *corpora* studies [6]. This kind of antonymy can be found in synonym and antonym dictionaries, as well. This research uses this last kind of search to find the direct antonyms.

2) *indirect* or *conceptual* – It does not occur between the word's forms, but between the word's senses. It is formed by semantic similarity with other adjectives, because some of them do not have direct antonyms, e.g. *obese*. Therefore, *obese* has an indirect antonym (*thin)* through its synonym *fat.*

[1] Gradable adjectives are those used with adverbs like *very, extremely, few, etc.*

2 Methodology

This research is based on the methodology proposed by [7] which divides the NLP studies into three complementary activities, according to three domains: 1) Linguistic; 2) Computational-Linguistic; 3) Computational.

Each domain presents some central problems and resources to solve these problems, as shown in Fig. 3:

DOMAINS	PROBLEMS	RESOURCES
Linguistic	Describe the linguistic knowledge and use	Linguistic Theories of Competence and Performance
↑ ↓	↑ ↓	↑ ↓
Computational-Linguistic	Represent the Linguist Domain knowledge	Formal Languages of Representation
↑ ↓	↑ ↓	↑ ↓
Computational	Code the Computational-Linguistic knowledge in a Programming Language	Programming Languages and Computer Systems

Fig. 3. The NLP domains, problems and resources (extracted from [7] p. 124)

In this research, those three domains may be represented by the following tasks:
1. Linguistic-related Domain: a) the study of the class of adjectives, with the investigation of their features and categories; b) the study of the antonymy, specifically for adjectives; c) the analysis of the BP adjectives selected from the corpus to specify the direct and indirect antonymy;
2. Linguistic-Computational Domain: a) the organization of the direct and indirect antonymy, following the model of WN.Pr, at TEP, in the adjectives collected from the corpus; b) the possible integration of the research results at TEP;
3. Computational Domain: it is not considered in this work.

The study is based on the occurrence of the 100 most frequent adjectives in a Brazilian Portuguese corpus, called *Mac-Morpho,* from *LacioWeb* project [8]. This corpus contains newspaper articles published in 1994, it has 1,167,183 words and it is morpho-syntactically annotated by *Palavras* parser.

One hundred adjectives were initially chosen to start the research, as it was necessary to establish a limit in their number. At the end of the research there was a list of 135 adjectives due to the direct and indirect antonymy.

Some tools were used to make the extraction of the 100 most frequent adjectives from the corpus: a) *Unitex* [9] as the concordancer used to find the adjectives and the sample sentences (contexts of use). These adjectives were extracted manually; b) WN.Pr for the antonymy representation model; c) some dictionaries of antonyms and

synonyms of BP [10] [11]; d) TeP as the source for the adjectives polisemy observation.

After the adjectives extraction, there were three basic steps in this work:

1. Establishment of the three adjectival categories: descriptive, relational and determinative. Some adjectives may be in two categories at the same time, according to the nouns they modify, e. g. 'econômico' (*economic)* is relational in 'crise econômica (=da economia)' (*economic (=of economy) crisis)* and is *descriptive* in 'carro econômico(=eficiente)' (*economic (=efficient) car);*
2. Direct antonymy formation, e.g. 'preto/branco' (*black/white)* with the aid of some dictionaries;
3. Indirect antonymy formation through: a) the polisemy of each adjective of the antonymic pair; b) the gradation of some adjectives, like 'fundamental=importante/insignificante' (*fundamental=important/insignificant).*

   The direct or indirect antonymy, the frequency number of each adjective and sample sentences were inserted in a table. Table 2 shows the example of 'velho' (*old):*

Table 2. The table of the adjective *old*

Headword	Synonym	Direct Antonym	Indirect Antonym	Frequency Number	Sample Sentence
velho (*old)*	idoso (*elder)*	jovem (*young)*		*142*	A mulher velha (*The old woman)*
	antigo (*antique)*	novo (*new)*	atual (*current)*		A igreja velha (*The old church)*

3 Results and Conclusion

From the steps shown above, it was possible to arrive at the following results:

1. There are 92 descriptive uses of adjectives among the 100 most frequent adjectives in the corpus;
2. There are 37 lexical pairs of antonyms, e.g. 'grande/pequeno' (*big/little)*. The other 19 pairs are formed with some prefix like: '*im'*- 'possível/impossível' (*possible/impossible), 'inter'* – 'nacional/internacional' (*national/international),* etc.;

3. Direct antonymy is more frequent among the analyzed adjectives than indirect antonymy. There are 80 adjectives with direct antonymy and just 12 with indirect antonymy;
4. There were 18 co-occurring adjective pairs in the corpus;
5. The prototypical relational adjectives are those related to country names. They don't have antonyms, e.g. 'brasileiro = do Brasil' (*Brazilian= from Brazil);*
6. In BP, some descriptive adjectives become determinative when they are placed before the noun, e.g. 'diferente' (*different)* in 'casas diferentes' e 'diferentes casas' ('*different houses').*

Concluding, this research aims at contributing to: a) Linguistics, mainly to Morphology studies through the adjectives analysis and to Semantics through the antonymy analysis; b) Computational Linguistics through the refinement of the antonymy representation at TeP.

References

1. FELLBAUM, C. (Ed.).: WordNet: an electronic lexical database. Cambridge, MA: MIT Press (1998).
2. DIAS-DA-SILVA, B. C. Wordnet.Br: An exercise of human language technology research. Revista Palavra, Série Linguagem, Processamento Automático do Português, N. 12, pp. 301-303 (2004).
3. MAZIERO, E.G.; PARDO, T.A.S.; DI FELIPPO, A.; DIAS-DA-SILVA, B.C. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. VI WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL), pp. 390-392 (2008). http://www.nilc.icmc.usp.br/tep2
4. MURPHY, M. L.: Semantic relations and the lexicon: antonymy, synonymy, and other paradigms. UK: Cambridge University Press (2003).
5. LYONS, J. Semantics.: Cambridge: Cambridge University Press, vol. 2 (1977).
6. JONES, S. Antonymy: a corpus – based approach. London: Routledge (2002).
7. DIAS-DA-SILVA, B.C. O estudo Linguístico-Computacional da Linguagem. Letras de Hoje, Vol. 41, N.2, pp. 103-138 (2006).
8. ALUÍSIO, S. M. et al.: An account of the challenge of tagging a reference corpus of Brazilian Portuguese. Relatório técnico NILC-TR-03-04. Fevereiro de 2003. http://www.nilc.icmc.usp.br/lacioweb/downloads/NILC-TR-03-04.zip
9. PAUMIER, S.: Unitex - manuel d'utilisation, research report. University of Marne-la-Vallée (2002).
10. BARBOSA, O.: Grande Dicionário de Sinônimos e Antônimos. Rio de Janeiro: Ediouro (1999).
11. FERNANDES, F.: Dicionário de Sinônimos e Antônimos da Língua Portuguesa. São Paulo: Globo (1997).