

Folheador: browsing through Portuguese semantic relations



Hugo Gonalo Oliveira & Hernani Costa & Diana Santos

{hroliv,hpcosta}@dei.uc.pt,
d.s.m.santos@ilos.uio.no

CISUC, University of Coimbra
FCCN, Linguateca
University of Oslo, Norway



Fundação para a Computação Científica Nacional

MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Introduction

Folheador^{ab} is an (open source^c) online browser for Portuguese Lexical Knowledge Bases (LKBs)

1. Navigation through Portuguese semantic relations, acquired from different sources
2. Linked to Portuguese corpora access providers
 - ▶ observation of related words co-occurring in authentic contexts

Current LKB

Semantic relations of CARTÃO [3], extracted from:

- ▶ Public domain electronic dictionaries
 - ▶ Dicionário da Língua Portuguesa, Porto Editora (see PAPEL [4])
 - ▶ Dicionário Aberto [8]
 - ▶ Wiktionary.PT^d
- ▶ Public domain thesauri
 - ▶ TeP [5]
 - ▶ OpenThesaurus.PT^e

Represented as relational triples x RELATED-TO y

- ▶ About 723k triples
 - ▶ synonymy, hypernymy, part-of, member-of, causation, producer-of, purpose-of, place-of, and property-of
- ▶ About 70k lexical items
 - ▶ 93,612 nouns, 38,409 verbs, 33,497 adjectives and 3,867 adverbs

Browsing features

- ▶ Search for all relations with one, two, or no fixed arguments, and one or no types
 - ▶ all lexical items related to a particular item
 - ▶ all relations between two lexical items
 - ▶ a sample of relations of a particular type
- ▶ Filter matching triples according to resource
- ▶ Provides the PoS of the lexical items
- ▶ Navigation through clicks on the items
- ▶ Queried item is always the first argument – relation is inverted on-the-fly
- ▶ Confidence values:
 - ▶ mere co-occurrence of the words in corpora
 - ▶ co-occurrence of the related words instantiating discriminating patterns of the particular relation
- ▶ (Partial) Graph visualisation

▶ In **Figure 1**, search for item *computador*:

- ▶ *computador* is usually a noun (e.g. in *computador* HIPONIMO_DE *máquina*)
- ▶ But might be an adjective (e.g. in *computador* PROPRIEDADE DO QUE *computar*)
- ▶ *computador* HIPONIMO_DE *aparelho* results of the inversion of *computador* HIPONIMO_DE *máquina*

The use of corpora

- ▶ Folheador is connected to AC/DC [7, 6]
 - ▶ access to Portuguese corpora
 - ▶ query for all the sentences connecting the arguments of a retrieved triple
 - ▶ e.g. in **Figure 2**, sentences with *computador* and *aparelho*
- ▶ Folheador is connected to VARRA [2]
 - ▶ based on patterns that express some of the relation types
 - ▶ occurrences of a triple in AC/DC
 - ▶ designed to allow human users to classify sentences according to whether they validate the relation, are just compatible with it, or not even that.
 - ▶ e.g. in **Figure 3**, sentences for (*computador* HIPONIMO_DE *máquina*)

Figures

The screenshot shows the Folheador search interface. At the top, there's a search bar with 'computador' entered. Below it, a table of results is displayed. The table has columns for 'TERMO1', 'RELAÇÃO', 'TERMO2', 'RECURSO(S)', and 'GRAU DE CONFIANÇA'. The results show various semantic relations like HIPONIMO_DE, HUPERONIMO_DE, and PROPRIEDADE DO QUE. The interface also includes a 'Pesquisar' button and a 'Última atualização' note.

Figure 1: Folheador's interface.

The screenshot shows a list of sentences from the AC/DC corpora. The sentences are related to the search term 'computador' and its related word 'aparelho'. The sentences are in Portuguese and illustrate various contexts where these words are used together.

Figure 2: AC/DC: some sentences returned for the related words *computador* and *aparelho*.

Relação	Procura	Exemplo
máquina HIPERONIMO_DE computador	padrões usados	par=Mais--94a-2: E também de ensinar máquinas como computadores a identificarem 'ses objetos . (NSC)
máquina HIPERONIMO_DE computador	padrões usados	par=ext328388-soc-95a-2: Máquinas como os computadores, os faxes e os videofones devem poder comunicar entre si sem falhas, o que supõe um trabalho de programação importante . (CP)

Figure 3: VARRA: sentences that exemplify the relation *computador* hyponym-of *máquina*.

Acknowledgements

Folheador was developed under the scope of Linguateca, throughout the years jointly funded by the Portuguese Government, the European Union (FEDER and FSE), UMIC, FCCN and FCT. Hugo Gonalo Oliveira is supported by the FCT grant SFRH/BD/44955/2008, co-funded by FSE.

References

- [1] Hernani Costa. O desenho do novo Folheador. Technical report, Linguateca, 2011.
- [2] Cláudia Freitas, Diana Santos, Hugo Gonalo Oliveira, and Violeta Quental. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In *Livro do IX Encontro de Linguística de Corpus*, ELC 2010, 2010.
- [3] Hugo Gonalo Oliveira, Leticia Antón Pérez, Hernani Costa, and Paulo Gomes. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários eletrônicos. *Linguamática*, 3(2):23–38, December 2011.
- [4] Hugo Gonalo Oliveira, Diana Santos, and Paulo Gomes. Extração de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, May 2010.
- [5] Erick G. Maziero, Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. Dias-da-Silva. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392, 2008.
- [6] Diana Santos. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3(2):113–128, June 2011. Volume edited by J.B.Johannessen, Language variation infrastructure.
- [7] Diana Santos and Eckhard Bick. Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of 2nd LREC*, pages 205–210, Athens, Greece, 2000.
- [8] Alberto Simões and Rita Farinha. Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, pages 159–171, December 2011.

^aFor using Folheador, go to <http://www.linguateca.pt/Folheador/>

^bSee technical details in [1]

^cSource code on <http://code.google.com/p/folheador/>

^d<http://pt.wiktionary.org>

^e<http://openthesaurus.caixamagica.pt>