Browsing Portuguese with Folheador

Hugo Gonçalo Oliveira[1], Hernani Costa[1,2], and Diana Santos[2,3]
hroliv@dei.uc.pt, hpcosta@dei.uc.pt, d.s.m.santos@ilos.uio.no

[1] CISUC, University of Coimbra, Portugal
[2] FCCN, Linguateca,
[3] University of Oslo, Norway

Lexical knowledge bases (LKBs) hold information on the words of a language and their interactions, according to their possible meanings. They are typically structured on word senses, which may be connected by semantic relations.

Regarding the complexity of most LKBs, their data formats are generally not suited for being read by humans. User interfaces have thus been developed for providing easier ways of exploring the LKB and assessing its contents. WordNet Search[4], for Princeton WordNet [2], or MNEX[5], for MindNet [9], are examples of such interfaces. However, in addition to information on words and semantic relations, it is important that the interfaces provide usage examples where semantic relations hold, or at least where related words co-occur.

Folheador [5] is a browser for Portuguese LKBs. Besides an interface for navigating through semantic relations acquired from different sources, it is linked to services that provide access to Portuguese corpora, thus allowing the observation of related words co-occurring in authentic contexts of use.

1 Browseable contents

Currently, Folheador browses through a LKB that integrates semantic triples obtained from: (i) PAPEL [6], a public domain lexical-semantic network; (ii) Portuguese handcrafted thesauri , TeP [1] and OpenThesaurus.PT[6]; (iii) Wiktionary.PT and Dicionário Aberto [8], both public domain dictionaries (extracted in the scope of project Onto.PT (more details in [4]).

Underlying relational triples are in the form x RELATED-TO y – one sense of lexical item x is related to one sense of lexical item y, by means of a relation identified by RELATED-TO.

2 Navigation

Folheador may be used for searching for all relations with one, two, or no fixed arguments, and one or no (predefined) types. Combining these options, we may obtain, for instance: all lexical items related to a particular item; all relations between two lexical items; or sample relations of a particular type.

[4] See http://wordnetweb.princeton.edu/perl/webwn
[5] See http://stratus.research.microsoft.com/mnex/
[6] See http://openthesaurus.caixamagica.pt/

Each matching triple is listed and may be filtered according to its source. The PoS of its arguments is shown and the arguments are links that ease navigation.

Figure 2 shows the result of searching for the word *computador*. In most of the retrieved triples, *computador* is a noun, but there are relations where it is an adjective. Even though we have not yet computed confidence values for all triples, when these values exist, Folheador presents, for each triple, a confidence value based on the co-occurrence of the arguments in corpora.

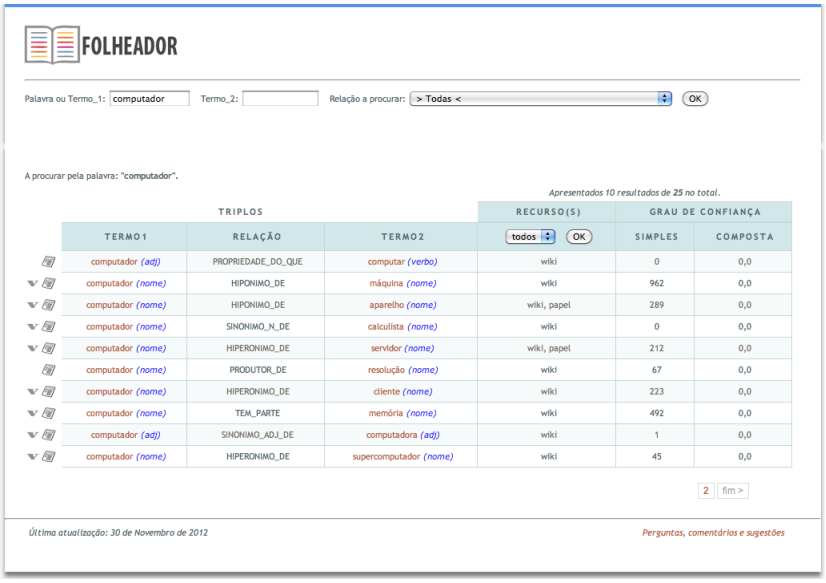


Fig. 1. Folheador's interface.

3 The use of corpora

Most lexical resources do not integrate frequency information and conflate highly specialized and obsolete words that co-occur with important and productive relations in everyday use. This is not good for human nor automatic users. Using corpora allows to add frequency information to both relation arguments and the triple, thus providing another axis to the description of words.

Also, it is interesting to observe language use in context, especially when the user is not sure whether the relation is correct or still in use. A corpus check provides illustration to a user facing an unusual or surprising relation, in addition to evaluation data for the relation curator or lexicographer.

Folheador is connected to AC/DC [7], an online service that provides access to Portuguese corpora, so that one can inspect all sentences in the AC/DC corpora that include both members of a particular triple, see figure 2 for the words *computador* and *aparelho*.

We also provide access to VARRA [3], another service for inspecting semantic relations in corpora through discriminating patterns for each relation. In Figure 3, we show two sentences returned for *computador* HIPONIMO_DE *máquina*.

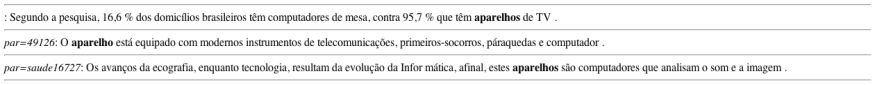
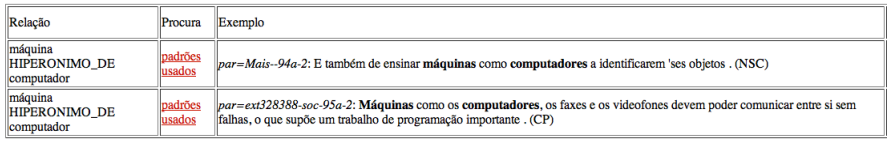


Fig. 2. Some sentences returned for the related words *computador* and *aparelho*.



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)

Fig. 3. 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 Gonçalo Oliveira is supported by the FCT grant SFRH/BD/44955/2008 co-funded by FSE.

References

1. Dias-Da-Silva, B.C., de Moraes, H.R.: A construção de um thesaurus eletrônico para o português do Brasil. ALFA 47(2), 101–115 (2003)
2. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
3. Freitas, C., Santos, D., Gonçalo Oliveira, H., Quental, V.: VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In: Livro do IX Encontro de Linguística de Corpus. ELC 2010 (2010)
4. Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., Gomes, P.: Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. Linguamática 3(2), 23–38 (December 2011)
5. Gonçalo Oliveira, H., Costa, H., Santos, D.: Folheador: browsing through Portuguese semantic relations. In: Procs. of 12th Conference of the European Chapter of the Association for Computational Linguistics (Demos Session). EACL 2012, ACL, Avignon, France (April 2012)
6. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. Linguamática 2(1), 77–93 (May 2010)
7. Santos, D., Bick, E.: Providing Internet access to Portuguese corpora: the AC/DC project. In: Procs. of 2nd LREC. pp. 205–210. Athens, Greece (2000)
8. Simões, A., Farinha, R.: Dicionário Aberto: Um novo recurso para PLN. Vice-Versa pp. 159–171 (December 2011)
9. Vanderwende, L., Kacmarcik, G., Suzuki, H., Menezes, A.: Mindnet: An automatically-created lexical resource. In: Procs. of HLT/EMNLP 2005 Interactive Demonstrations. pp. 8–9. ACL, Vancouver, Canada (2005)