

PAPEL: a lexical ontology for Portuguese

Hugo Gonçalo Oliveira, Paulo Gomes, Nuno Seco

Linguateca, node of Coimbra, DEI-FCTUC, CISUC

Diana Santos

Linguateca, node of Oslo, SINTEF ICT

*Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática*



*Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra*





Contents

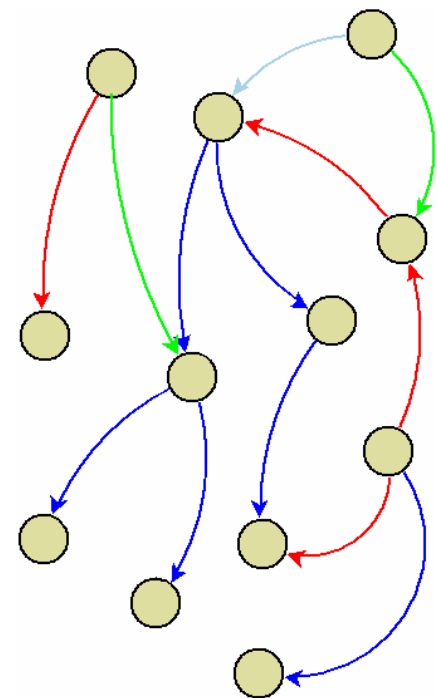
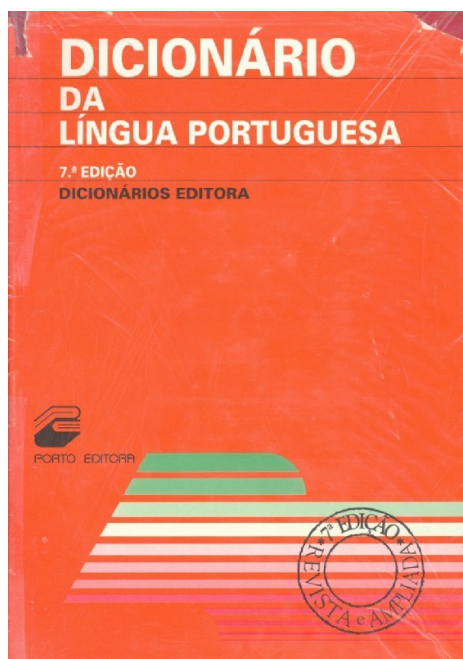


- **Project objectives**
- **MRDs (as a source of semantic knowledge)**
- **Similar resources**
- **Building PAPEL**
- **Detailed example**
- **Current precision**
- **Further work/possible directions**

Project objectives



PAPEL stands for Palavras Associadas Porto Editora Linguateca



MRDS as a source of semantic knowledge



- Since 1970's (Calzolari 1977, Amsler 1981, ...)
- Restricted and predictable vocabulary
- Typical definition structure:
genus + differentia
 - *genus*: the superordinate concept
 - *differentia*: properties for distinction between instances of the same superordinate concept.

Similar resources



Princeton WordNet (Miller 1990)

- freely available
- widely used by NLP researchers
- created “from scratch”
- synsets + relations
- different kinds of relations for words of different grammatical categories
- Hypernymy/Hyponymy, Meronymy/Holonymy.

Similar resources



Microsoft's MindNet (Richardson et al. 1998)

- automatically created with tools to extract relations from MRDs
- a large amount of relations: Attributive modification, Causation, Classifier, Goal, Hypernymy, Intensifier, Location, Manner, Means...
- a web interface: MNEX
(<http://atom.research.microsoft.com/mnex/>)

Similar resources



FrameNet (Baker et al. 1998)

- description of semantic frames

WordNet.BR (Dias da Silva 2004)

- as far as we know only contains the Synonymy and Antonymy relations.

WordNet.PT (Marrafa et al. 2006)

- little information shared
- website/search interface has not been working for the past year...

Building PAPEL



Examples of relations to include:

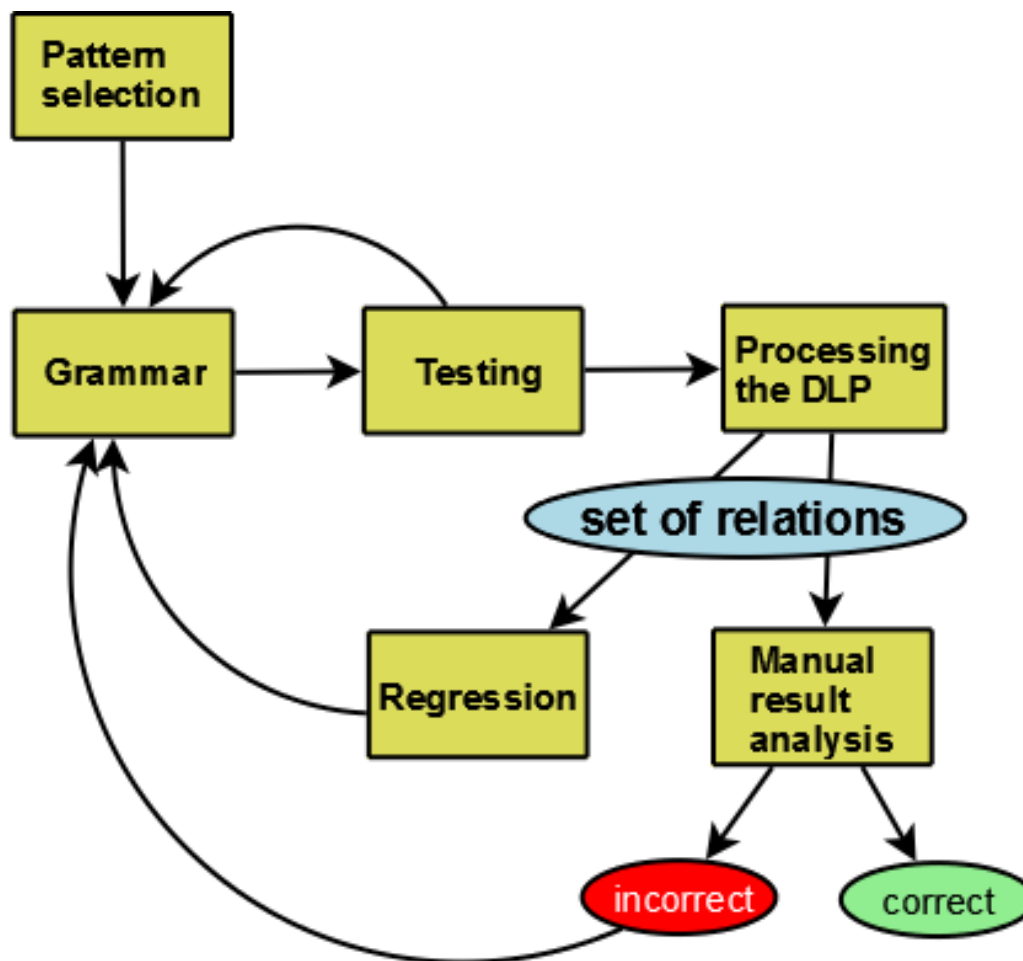
Relation	Read	Example
HIPERONIMO-DE(X, Y)	X is a HIPERONIMO-DE Y	HIPERONIMO-DE(animal, cão)
HIPONIMO-DE(X, Y)	X is a HIPONIMO-DE Y	HIPONIMO(cão, animal)
CAUSADOR-DE(X, Y)	X is a CAUSADOR-DE Y	CAUSADOR-DE(vírus, doença)
RESULTADO-DE(X, Y)	X is a RESULTADO-DE Y	RESULTADO-DE(doença, vírus)
MEIO-PARA(X, Y)	X is a MEIO-PARA Y	MEIO-PARA(chave, abrir)
FINALIDADE-DE(X, Y)	X is a FINALIDADE-DE Y	FINALIDADE-DE(abrir, chave)
PARTE-DE(X, Y)	X is a PARTE-DE Y	PARTE-DE(roda, carro)
INCLUI(X, Y)	X INCLUI Y	INCLUI(carro, roda)
LOCAL-DE(X, Y)	X is a LOCAL-DE Y	LOCAL-DE(restaurante, comer)
OCORRE-EM(X, Y)	X OCORRE-EM Y	LOCALIZADO-EM(comer, restaurante)

Building PAPEL



- **Quantitative studies about the patterns used in the definitions.**
- **Using the chart parser PEN with specific grammars to extract relations.**
- **Semantic relations from definitions:**
 - `Y - um tipo de X → HIPERONIMO(X, Y)`

Building PAPEL



Detailed example



Causation relation:

- relation between an agent (the causer) and a result (the caused)
- CAUSADOR-DE and its inverse RESULTADO-DE.
- Based on verbs: *causar, originar, provocar, produzir, gerar, motivar, suscitar.*

Detailed example



Examples of (simplified) rules for CAUSADOR-DE:

- {causad|originad|provocad|produzid|gerad|motivad|suscitad}{o|os|a|as} **FREQ* PREP CAUSADOR**
 - concussão (s.f.) - choque violento **originado por** uma explosão
 - **CAUSADOR-DE**(explosão, concussão)
- efeito **PREP CAUSADOR**
 - maximização (s.f) - **efeito de** maximizar
 - **CAUSADOR-DE**(maximizar, maximização)
- devido {a|ao|à|às|aos} **CAUSADOR**
 - engasgo (s.m.) - incapacidade de respirar **devido a** obstrução da garganta
 - **CAUSADOR-DE**(obstrução , engasgo)

Detailed example



Examples of (simplified) rules for RESULTADO-DE:

- que {causa|origina|provoca|produz|motiva|gera|suscita}
RESULTADO
 - `diplodoco (s.m.) - bactéria que causa as meningites cerebrospinais`
 - **RESULTADO-DE**(meningites, diplodoco)
- {causar|originar|provocar|produzir|motivar|gerar|sus citar}
RESULTADO
 - `incomodidade (s.f.) - o que causa mal-estar ou desconforto`
 - **RESULTADO-DE**(mal-estar, incomodidade)
 - **RESULTADO-DE**(desconforto, incomodidade)

Detailed example



Examples of current problems:

- Specific relations inside the definition
 - estetoscópio (s.m) - instrumento para auscultar a respiração, as batidas do coração e outros sons produzidos pelo corpo.
 - CAUSADOR-DE(corpo, estetoscópio)
- Negation of patterns
 - respeitar (v. tr.) - não **causar** dano
 - RESULTADO-DE(dano, respeitar)



Current precision



Relation name	Runs	Hits	Tagged	Tagged correct	Tagged incorrect	Precision
CAUSADOR-DE	4	5657	3354	3222	132	0,96
RESULTADO-DE	4	1693	1070	972	98	0,91
FINALIDADE-DE	1	1348	192	184	8	0,96
MEIO-PARA	1	826	286	258	28	0,9
INCLUI	2	1908	328	320	8	0,98
PARTE-DE	2	811	202	171	31	0,85

Number of definitions in the whole DLP = 237,246

Further work



- **Improvements to PEN**
- **Using a broad-coverage parser like PALAVRAS (Bick 2000)**
- **Validation of semantic relations by people.**
- **Construction of the network:**
 - grouping the words into synsets
 - word sense desambiguation

The end...



Questions/suggestions?

Acknowledgement

- *This work was done in the scope of the Linguateca, contract n°339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.*





References



- Nicoletta Calzolari. An empirical approach to circularity in dictionary definitions. In *Cahiers de Lexicologie*, pages 118-128, 1977.
- Robert A. Amsler. A taxonomy for english nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pages 133-138, Morristown, NJ, USA, 1981. Association for Computational Linguistics.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database*. *Int J Lexicography*, 3(4):235244, January 1990.
- Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. Mindnet: acquiring and structuring semantic information from text. In *Proceedings of the 17th international conference on Computational linguistics*, pages 10981102, Morristown, NJ, USA, 1998. Association for Computational Linguistics.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 8690, Morristown, NJ, USA, 1998. Association for Computational Linguistics.
- Bento C. Dias da Silva. Wordnet.br: An exercise of human language technology research. *Revista PaLavra*, no. 12, 2004. Série Linguagem, Processamento Automático do Português:1524, 2004.
- P. Marrafa, R. Amaro, R. P. Chaves, S. Lourosa, C. Martins, and S. Mendes. Wordnet.pt new directions. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (eds.), *Proceedings of GWC'06: 3rd International Wordnet Conference*, Jeju Island, Korea, pages 319320, 2006.
- Eckhard Bick. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University, Arhus, 2000.