

Relations extracted from a Portuguese dictionary: results and first evaluation

Hugo Gonalo Oliveira*, Diana Santos, Paulo Gomes
hroliv@dei.uc.pt, diana.santos@sintef.no, pgomes@dei.uc.pt

* Hugo Gonalo Oliveira is supported by FCT, grant SFRH/BD/44955/2008.



*University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering*

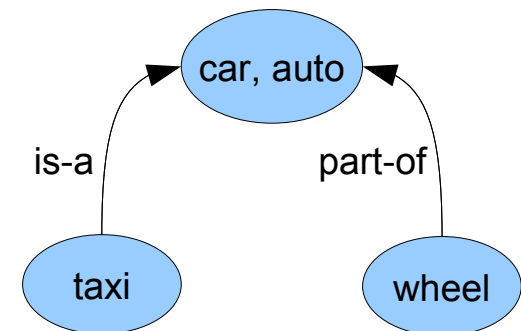


*Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra*



Introduction

- Natural language processing needs access to semantic knowledge
- Lexical ontologies
 - Conceptual models of a language, structured on words and meanings
 - Lexico-semantic relations
 - Synonymy: *car* syn *auto*
 - Hyponymy (is-a): *taxi* is-a *car*
 - Meronymy (part-of): *wheel* part-of *car*
 - Cause, purpose, location...



Construction of an ontology

■ Handcraft:

- More reliable
- Difficult to maintain and update

■ Semi-automatic:

- From dictionaries
 - Semantic authorities, restricted vocabulary
 - General knowledge
- From corpora
 - Much available, rich on specific domains
 - Unrestricted text, harder to process

Lexical ontologies

For English

■ Handcrafted:

- Princeton WordNet

- Cyc

- Berkeley FrameNet

■ Semi-automatic construction:

- MindNet

Lexical ontologies

For Portuguese

■ Handcrafted:

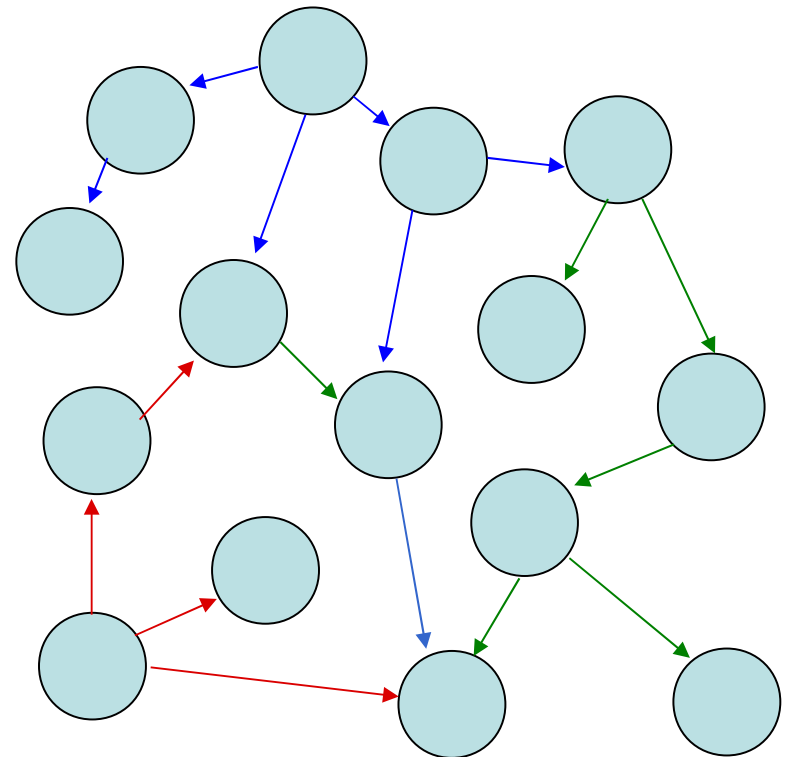
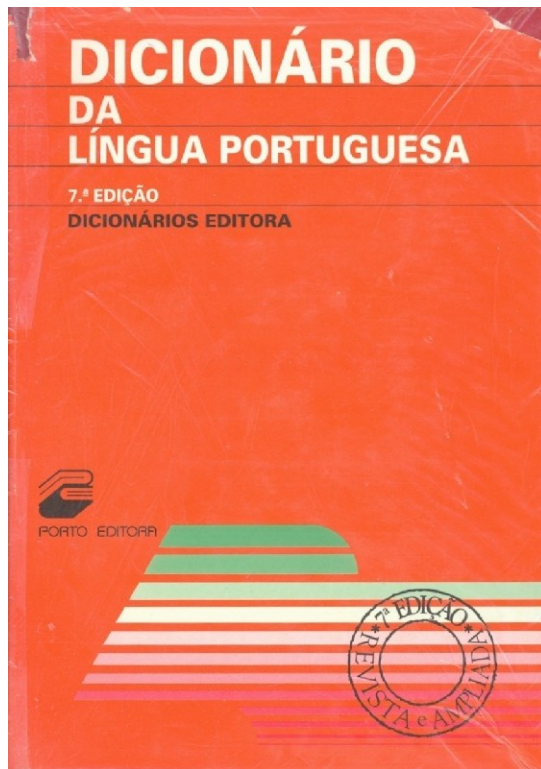
- Tep (<http://www.nilc.icmc.usp.br/tep2/>)
- WordNet.BR (<http://www.nilc.icmc.usp.br/~arianidf/WordNet-BR.html>)
- WordNet.PT (<http://cvc.instituto-camoes.pt/wordnet/>)
- MultiWordNet.PT (<http://mwnpt.di.fc.ul.pt/>)

■ Semi-automatic construction:

- PAPEL (<http://www.linguateca.pt/PAPEL>)

PAPEL

■ Palavras Associadas Porto Editora Linguateca



PAPEL

- About 200,000 relational triples between Portuguese terms, in PAPEL 1.1:

Group	Name	Args.	Qnt.	Examples
Synonymy	SINONIMO_N_DE	n,n	37,259	(<i>auxílio, contributo</i>)
	SINONIMO_V_DE	v,v	21,534	(<i>tributar, colectar</i>)
	SINONIMO_ADJ_DE	adj,adj	19,073	(<i>flexível, moldável</i>)
	SINONIMO_ADV_DE	adv,adv	1,169	(<i>após, seguidamente</i>)
Hypernymy	HIPERONIMO_DE	n,n	61,477	(<i>planta, salva</i>)
Meronymy	PARTE_DE	n,n	9,970	(<i>cauda, cometa</i>)
	PARTE_DE_ALGO_COM_PROP	n,adj	3,806	(<i>tampa, coberto</i>)
	PROP_DE_ALGO_PARTE_DE	adj,n	900	(<i>celular, célula</i>)
Cause	CAUSADOR_DE	n,n	1,010	(<i>fricção, assadura</i>)
	CAUSADOR_DE_ALGO_COM_PROP	n,adj	17	(<i>paixão, passional</i>)
	PROP_DE_ALGO_CAUSADOR_DE	adj,n	498	(<i>reactivo, reacção</i>)
	ACCAO_QUE_CAUSA	v,n	6,399	(<i>limpar, purgação</i>)
	CAUSADOR_DA_ACCAO	n,v	39	(<i>gases, fumigar</i>)
Producer	PRODUTOR_DE	n,n	885	(<i>romãzeira, romã</i>)
	PRODUTOR_DE_ALGO_COM_PROP	n,adj	34	(<i>sublimação, sublimado</i>)
	PROP_DE_ALGO_PRODUTOR_DE	adj,n	359	(<i>fotógeno, luz</i>)
Purpose	FINALIDADE_DE	n,n	2,878	(<i>defesa, armadura</i>)
	FINALIDADE_DE_ALGO_COM_PROP	n,adj	38	(<i>reprodução, reproduutor</i>)
	ACCAO_FINALIDADE_DE	v,n	5,185	(<i>fazer_rir, comédia</i>)
	ACC_FINALIDADE_DE_ALGO_COM_PROP	v,adj	284	(<i>corrigir, correccional</i>)
Place	LOCAL_ORIGEM_DE	n,n	816	(<i>Japão, japones</i>)
Manner	MANEIRA_POR_MEIO_DE	adv,n	1,113	(<i>timidamente, timidez</i>)
	MANEIRA_SEM	adv,n	121	(<i>devagar, pressa</i>)
	MANEIRA_SEM_ACCAO	adv,v	11	(<i>assiduamente, faltar</i>)
Property	PROP_DE_ALGO_REFERENTE_A	adj,n	3,520	(<i>dinâmico, movimento</i>)
	PROP_DO_QUE	adj,v	17,246	(<i>familiar, ser_conhecido</i>)

- General relation names (HIPERONIMIA, PARTE, CAUSA ...)
- More specific names (and inverse), according to the arguments grammatical categories

```
PARTE {
nome:nome * PARTE_DE:INCLUI;
nome:adj * PARTE_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_QUE_INCLUI;
adj:nome * PROPRIEDADE_DE_ALGO_PARTE_DE:INCLUI_ALGO_COM_PROPRIEDADE;
}

CAUSA{
nome:nome * CAUSADOR_DE:RESULTADO_DE;
nome:verbo * CAUSADOR_DA_ACCAO:ACCAO_RESULTADO_DE;
nome:adj * CAUSADOR_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_RESULTADO_DE;
adj:nome * PROPRIEDADE_DE_ALGO_QUE_CAUSA:RESULTADO_DE_ALGO_COM_PROPRIEDADE;
verbo:nome * ACCAO_QUE_CAUSA:RESULTADO_DA_ACCAO;
}
```

■ PEN* parser + semantic grammars

1

cometa, s. m.

astro geralmente constituído por núcleo, cabeleira e cauda

3

núcleo PARTE_DE cometa
cabeleira PARTE_DE cometa
cauda PARTE_DE cometa

2

```
[RAIZ]
  [QUALQUERCOISA]
    > [astro]
      [QUALQUERCOISA]
        > [geralmente]
          [PADRAO_CONSTITUIDO]
            [VERBO_PARTE_PP]
              > [constituído]
                [PREP]
                  > [por]
                    [ENUM_PARTE]
                      [PARTE_DE]
                        > [núcleo]
                          [VIRG]
                            > [,]
                              [ENUM_PARTE]
                                [PARTE_DE]
                                  > [cabeleira]
                                    [CONJ]
                                      > [e]
                                        [PARTE_DE]
                                          > [cauda]
```

*available through <http://code.google.com/p/pen/>

PAPEL

More examples

barranquenho, s. m. natural ou habitante de **Barrancos**

--> Barrancos LOCAL_ORIGEM_DE barranquenho

clarabóia, s. f. **janela** ou **fresta** por onde entra a luz num aposento

--> fresta HIPERONIMO_DE clarabóia

--> janela HIPERONIMO_DE clarabóia

vínculo, s. m. tudo o que serve para **prender** ou **atar**

--> prender ACCAO_FINALIDADE_DE vínculo

--> atar ACCAO_FINALIDADE_DE vínculo

cólera, s. f. grave doença epidémica, contagiosa, que provoca **diarreia**, **vómitos** e **cólicas**, causada por um **bacilo** (vibrião) e também designada por cólera-morbo e cólera-asiática, mordexim ou mordixim

--> diarreia RESULTADO_DE cólera

--> vómitos RESULTADO_DE cólera

--> cólicas RESULTADO_DE cólera

--> bacilo CAUSADOR_DE cólera

- All relations converted to the direct type
 - *manga* INCLUI *punho* >> *punho* PARTE_DE *manga*
 - *dor* RESULTADO_DE *distensão* >>
distensão CAUSADOR_DE *dor*
- Lemmatization of arguments
- Relation's name correction
 - *loucura* ACCAO_QUE_CAUSA *desvario* >>
loucura CAUSADOR_DE *desvario*

Evaluation

Synonymy

- Tep 2.0 as a golden resource
- Relations with terms that are not both in PAPEL 1.0 and Tep 2.0 are removed
 - 50% of PAPEL in TeP, 39% of TeP in PAPEL
- Expansion: (A SINONIMO_DE B) e (B SINONIMO_DE C) >> (A SINONIMO_DE C)
 - 19% of PAPEL in TeP, 90% of TeP in PAPEL
 - Incorrections
 - A=*ruína*, B=*queda*, C=*habilidade*
 - >> *ruína SINONIMO_DE habilidade*

Evaluation

Other relations

- Relations rendered to natural language patterns
- Searched on CETEMPúblico, via AC/DC

Relação	Certa?	Justificação
<i>língua</i> HIPERONIMO_DE <i>italiano</i>	Sim	<i>As línguas latinas, como o italiano ou o português, tornam-se mais fáceis por causa das vogais.</i>
<i>arbusto</i> PARTE_DE <i>floresta</i>	Sim	<i>A floresta é um conjunto de árvores, arbustos e ervas de várias qualidades e tamanhos.</i>
<i>cólera</i> CAUSADOR_DE <i>diarreia</i>	Sim	<i>A cólera provoca fortes diarreias e vômitos e pode levar à desidratação e, conseqüentemente, à morte em poucas horas.</i>
<i>oliveira</i> PRODUTOR_DE <i>azeitona</i>	Sim	<i>Também a quantidade e tamanho das azeitonas produzidas por uma oliveira biológica é inferior, já que não são utilizados compostos de azoto que ajudam a planta a crescer.</i>
<i>recrutamento</i> FINALIDADE_DE <i>inspecção</i>	Sim	<i>Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às inspecções para recrutamento, revelou o ministro da Defesa.</i>
<i>músico</i> PARTE_DE <i>música</i>	Não	<i>... um espectáculo baseado na obra "Cantos de Maldoror", de Lautréamont, com música composta pelo músico inglês Steven Severin...</i>
<i>fim</i> FINALIDADE_DE <i>sempre</i>	Não	<i>Sicilia aponta sempre para o fim do dia, para o fim da luz.</i>

First evaluation results

■ Results for PAPEL 1.0 + CETEMPúblico 4.0

Relação	Relações c/ args no CETEMPúblico	%	Amostra	%	Encontradas	%
Hiperonímia	40,079	63%	3,145	8%	560	18%
Meronímia	3,746	35%	2,343	63%	521	22%
Causa	557	50%	557	100%	20	4%
Produtor	414	44%	414	100%	12	3%
Finalidade	1,718	59%	1,718	100%	173	10%

■ Some correct relations are not found:

- fruto HIPERONIMO_DE alperce
- algoritmia PARTE_DE matemática
- ausência CAUSADOR_DE saude
- aquecimento FINALIDADE_DE salamandra
- ...

Final remarks

- Starting point for future projects
 - New evaluations
 - Comparison with other resources
 - Integration with other resources and methodologies
 - Usage to query corpora (through AC/DC)
- PAPEL is completely free: everybody can do what they want with it, improve it, and redistribute it!
 - There is a lot to improve
 - New versions will be available during the next months

Soon...

- Diana Santos, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros & Rosário Silva. "O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o o PAPEL". In *XXV Encontro Nacional da Associação Portuguesa de Linguística* (Lisboa, Portugal, 22-24 de Outubro de 2009).

Acknowledgements

- PAPEL was developed in the scope of Linguateca, under contract POSC/339/1.3/C/NAC, co-funded by the Portuguese Government, by the European Union (FEDER and FSE), by UMIC e by FCCN.



- Hugo Gonçalo Oliveira is supported by FCT, scholarship grant SFRH/BD/44955/2008.
- We would also like to thank to the R&D group of Porto Editora

Relations extracted from a Portuguese dictionary: results and first evaluation

Thank you!

Hugo Gonalo Oliveira, Diana Santos, Paulo Gomes
hroliv@dei.uc.pt, diana.santos@sintef.no, pgomes@dei.uc.pt



*University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering*



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

