

## Avaliação da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário

Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes  
[hroliv@dei.uc.pt](mailto:hroliv@dei.uc.pt), [diana.santos@sintef.no](mailto:diana.santos@sintef.no), [pgomes@dei.uc.pt](mailto:pgomes@dei.uc.pt)

STIL 2009



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



## Ontologias lexicais

- Conjunto de termos e conceitos estruturados e ligados de acordo com relações léxico-semânticas
- Úteis para...
  - Interpretar textos, determinar semelhanças entre conceitos, resposta automática a perguntas, tradução automática, geração de texto, pesquisa inteligente, estudos teóricos acerca da semântica de uma língua...

Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes STIL 2009

## Ontologias lexicais

### Construção

- Manual
  - Menos propícia a erros
  - Difícil manutenção/actualização
- Semi-automática
  - A partir de dicionários
    - "autoridades" semânticas, vocabulário restrito
    - conhecimento geral
  - A partir de corpos
    - muita quantidade, rico em domínios específicos
    - texto muito variado e mais difícil de processar, não procuram abranger toda a língua

Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes STIL 2009

## Ontologias lexicais

### Exemplos

- Construção manual
  - Wordnet
  - Cyc
  - FrameNet
- Construção semi-automática
  - MindNet

Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes STIL 2009

## Ontologias lexicais

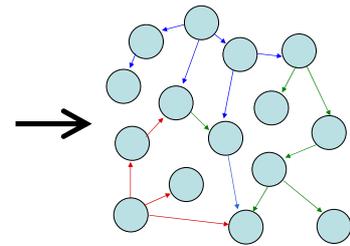
### Língua Portuguesa

- Construção manual
  - 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/>)
- Construção semi-automática
  - PAPEL (<http://www.linguateca.pt/PAPEL/>)

Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes STIL 2009

## PAPEL

- Palavras Associadas Porto Editora Linguateca



Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes STIL 2009

# PAPEL

- Conjunto de (cerca de 200 mil) tripos relacionais entre termos portugueses

Grupo	Nome	arg1_arg2	Núm.	Exemplos
Sinonímia	SINONIMO.DE	qpl_sarg1	80412	(gloriet, modêstev)
Hiperonímia	HIPERONIMO.DE	sub_sub	63455	(planta, salva)
Meronímia	PARTE.DE	sub_sub	14453	(comida, cometa)
	PARTE.DE_ALGO.COM_PROPRIEDADE	sub_adj	3715	(sampa, coelhos)
Causa	PROPRIEDADE.DE_ALGO.PARTE.DE	adj_sub	962	(célula, célula)
	CAUSADOR.DE	sub_sub	1125	(fricção, assadura)
	CAUSADOR.DE_ALGO.COM_PROPRIEDADE	sub_adj	16	(paído, passional)
	PROPRIEDADE.DE_ALGO.CAUSADOR.DE	adj_sub	515	(treceiro, reuço)
Produtor	ACCAO.QUE_CAUSA	v_sub	6424	(limpar, purgação)
	CAUSADOR.DA_ACCAO	sub_v	39	(gases, fumigar)
	PRODUTOR.DE	sub_sub	932	(romãlzeira, rima)
Fim	PRODUTOR.DE_ALGO.COM_PROPRIEDADE	sub_adj	31	(sublimação, sublimado)
	PROPRIEDADE.DE_ALGO.PRODUTOR.DE	adj_sub	348	(falsetro, fac)
	FINALIDADE.DE	sub_sub	2095	(passagem a nível, agregação)
	FINALIDADE.DE_ALGO.COM_PROPRIEDADE	sub_adj	23	(enumeração, enumerativo)
Lugar	ACCAO.FINALIDADE.DE	v_sub	5640	(ficar, vir, comêda)
	ACCAO.FINALIDADE.DE_ALGO.COM_PROPRIEDADE	v_adj	255	(corrigir, correcção)
Propriedade	MANEIRA.FOR.MODO.DE	adv_sub	1433	(simultaneamente, rímide)
	LOCAL.ORIGEM.DE	sub_sub	768	(lepto, japonês)
Propriedade	PROPRIEDADE.DE_ALGO.REFERENTE.A	adj_sub	3700	(diplômico, movimento)
	PROPRIEDADE.DO.QUE	adj_v	17028	(diplomado, possuir, diploma)

# PAPEL

- Extraídos de forma semi-automática, com base em gramáticas semânticas (específicas para cada relação) que incluem padrões indicadores

Padrão	Relação associada
tipo género classe forma de	Hiperonímia
parte membro de	Meronímia
que causa provoca origina	Causa
usado utilizado para	Objectivo
uma palavra ou lista de palavras	Sinonímia

# PAPEL

- Relações genéricas (REL\_GEN) e também mais específicas (REL\_ESP), com base nas categorias gramaticais dos argumentos, pré-definidas da seguinte forma:

```
REL_GEN {
  cat1:cat2 * REL_ESP_1+REL_ESP_1_inv;
  cat1:cat3 * REL_ESP_2+REL_ESP_2_inv;
}
```

```
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;
}
```

# PAPEL

- Extracção

- cometa, s. m.  
astro geralmente constituído por núcleo, cabeleira e cauda
- 
- núcleo PARTE\_DE cometa  
cabeleira PARTE\_DE cometa  
cauda PARTE\_DE cometa

```
[RAIZ]
[QUALQUERCOISA]
> [ARTIC]
[QUAIQUERCOISA]
> [qualmente]
[FADRAO_CONSTITUIDO]
[VERBO_PARTE_DE]
> [constituído]
[PREP]
> [por]
[ENUM_PARTE]
[PARTE_DE]
> [núcleo]
[VERB]
> [ ]
[ENUM_PARTE]
[PARTE_DE]
> [cabeleira]
[COM]
> [e]
[PARTE_DE]
> [cauda]
```

# PAPEL

## Mais exemplos de extracção

**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

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

- > fresta HIPERONIMO\_DE clarabóia
- > janela HIPERONIMO\_DE clarabóia

**tapadura**, s. f. objecto próprio para **tapar** ou **cobrir**

- > tapar ACCAO\_FINALIDADE\_DE tapadura
- > cobrir ACCAO\_FINALIDADE\_DE tapadura

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

- > prender ACCAO\_FINALIDADE\_DE vínculo
- > atar ACCAO\_FINALIDADE\_DE vínculo

# PAPEL

- Ajuste de relações:

- Tudo para tipo directo
  - manga** INCLUI **punho** >> **punho** PARTE\_DE **manga**
  - dor** RESULTADO\_DE **distensão** >> **distensão** CAUSADOR\_DE **dor**
- Lematização dos argumentos
- Correcção do nome das relações
  - loucura** ACCAO\_QUE\_CAUSA **desvario** >> **loucura** CAUSADOR\_DE **desvario**

## Avaliação

### Sinonímia

- Utilização do TeP como recurso dourado
- Remoção das relações com termos que não são comuns a ambos os recursos
  - 50% do PAPEL no TeP, 39% do TeP no PAPEL
- Expansão: (A SINONIMO\_DE B) e (B SINONIMO\_DE C) >> (A SINONIMO\_DE C)
  - 19% do PAPEL no TeP, 90% do TeP no PAPEL
- Incoerências
  - A=ruína, B=queda, C=habilidade
  - >> ruína SINONIMO\_DE habilidade

## Avaliação

### Restantes relações

- Avaliação das restantes relações
  1. Tradução das relações para "linguagem natural"
  2. Busca no CETEMPúblico, através do AC/DC

Relação	Certa?	Justificação
língua HIPERONIMO_DE italiano	Sim	As línguas latinas, como o italiano ou o português, tornam-se mais fáceis por causa dos vogais.
arbusto PARTE_DE floresta	Sim	A floresta é um conjunto de árvores, arbustos e ervas de várias qualidades e tamanhos.
cólera CAUSADOR_DE diarreia	Sim	A cólera provoca fortes diarreias e vômitos e pode levar à desidratação e, conseqüentemente, à morte em poucas horas.
oliveira PRODUTOR_DE azeitona	Sim	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.
recreamento FINALIDADE_DE inspeção	Sim	Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às inspeções para recreamento, revelou o ministro da Defesa
músico PARTE_DE música	Não	... um espectáculo baseado na obra "Canos de Maldoror", de Lauréramon, com música composta pelo músico inglês Steven Severin...
fim FINALIDADE_DE sempre	Não	Sicília aponta sempre para o fim do dia, para o fim da luz.

## Avaliação

### Usando o AC/DC

- Remoção das relações com argumentos inexistentes no CETEMPúblico, a partir das listas de frequência dos lemas
- Criação de padrões simples de interrogação:

```
PRED_CAUSA_A ::=
[ lema="poder|ir|ter|dever" ] *
[ lema="causar|provocar|originar|suscitar|motivar|resultar|produzir|gerar" ]
[ pos="ADV. * " ] *
INSTANCIA_DE ::=
[ lema="tipo|forma|instância|género" ]
[ lema="de " ]
```

## Resultados da primeira avaliação

### Apenas com padrões simples no CETEMPúblico v. 4.0

- Primeiros resultados

Relação	Relações e/ 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%

- Algumas relações correctas, mas não encontradas:
  - fruto HIPERONIMO\_DE alperce
  - algoritmia PARTE\_DE matemática
  - ausência CAUSADOR\_DE saude
  - aquecimento FINALIDADE\_DE salamandra
  - ...

## Resultados da primeira avaliação

### Apenas com padrões simples no CETEMPúblico v. 4.0 (cont.)

- Muitas das relações extraídas, por vezes até do foro etimológico, não aparecem naturalmente em texto noticioso
  - China LOCAL\_ORIGEM\_DE chinês
  - forma\_de\_célula PARTE\_DE\_ALGO\_COM\_PROPRIEDADE celuliforme
- Os padrões não cobrem por enquanto muitos casos
  - viga PARTE\_DE ponte
  - par=ext423909-soc-92a-1: Ferreira do Amaral conta poder adjudicar o reforço da ponte (com tirantes e uma nova viga) e a construção por privados da nova ferrovia dentro de ano e meio
- Vários falsos positivos, sobretudo nos casos de padrões muito genéricos

## Comentários finais

- Ponto de partida para projectos futuros
  - Novas avaliações
  - Adição de informação de fiabilidade às relações
  - Estudo quantitativo da polissemia
  - Comparação com outros recursos
  - Integração com outros recursos e outras metodologias
  - Seu uso para interrogação de corpos (AC/DC)
- O PAPEL é completamente livre: todos podem fazer o que quiserem com ele, melhorá-lo, e redistribuí-lo
  - Há muito para melhorar
  - Contamos com novas versões e avaliações em colaboração

## Agradecimentos

- O PAPEL foi desenvolvido no âmbito da Linguateca, sob o contrato POSC/339/1.3/C/NAC, co-financiado pelo governo português, pela União Europeia (FEDER e FSE), pela UMIC e pela FCCN.



- Hugo Gonçalo Oliveira é financiado pela FCT, bolsa SFRH/BD/44955/2008.
- Agradecemos ainda ao grupo de I&D da Porto Editora o apoio na construção do PAPEL.

## Muito em breve...

- 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 PAPEL". In *XXV Encontro Nacional da Associação Portuguesa de Linguística* (Lisboa, Portugal, 22-24 de Outubro de 2009).