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

Hugo Gonçalo Oliveira[1], **Diana Santos**[2], **Paulo Gomes**[1]

hroliv@dei.uc.pt, diana.santos@sintef.no, pgomes@dei.uc.pt

[1]CISUC, Universidade de Coimbra, Portugal

[2]Linguateca, SINTEF ICT, Noruega

Abstract. *This paper presents PAPEL, a lexical resource for Portuguese, consisting of relations between terms, extracted by (semi) automatic means from a general dictionary. After a short overview of the building process, a quantitative overview is given together with some examples. Evaluation is then presented and discussed: for synonymy, we used a public thesaurus, Tep, for the other relations, we queried Portuguese corpora through the AC/DC interface.*

Resumo. *Apresentamos o PAPEL, um recurso lexical para o português, constituído por relações entre termos, extraídas de forma (semi) automática de um dicionário da língua geral, fornecendo uma visão do processo de construção do recurso, e detalhando as relações incluídas e o seu número. Relatamos também a sua primeira avaliação, que tomou duas formas: para as relações de sinonímia, a comparação com um tesauro publicamente acessível; para as outras relações, interrogando corpos em português.*

1. Apresentação

Cada vez mais os estudos do processamento da língua exigem que haja acesso computacional a informação semântica, e é cada vez mais frequente o recurso a redes ou ontologias lexicais que tentam cobrir o panorama lexical de uma língua toda, ao invés, ou como complemento, de terminologias, cujo objectivo é descrever uma área específica do conhecimento. A ontologia paradigmática, para o inglês, é a WordNet [Fellbaum 1998], embora uma ontologia mais relacionada com o nosso trabalho seja a MindNet [Richardson et al. 1998].

Para o português, temos conhecimento de projectos de criação de uma ontologia lexical para a língua toda, abarcando apenas a variante brasileira [Dias da Silva et al. 2002], relativa ao português de Portugal [Marrafa 2002], e mais recentemente a ligação a iniciativas europeias como o MultiWordNet (`http://mwnpt.di.fc.ul.pt`). Estes projectos, conforme pensamos saber, seguem uma metodologia semelhante à usada na construção da WordNet de Princeton, ou seja, usam peritos para criar o recurso manualmente.

No que se refere a ontologias criadas directamente a partir de dicionários, pensamos que o projecto que apresentamos aqui – o PAPEL, Palavras Associadas Porto Editora - Linguateca (`http://www.linguateca.pt/PAPEL`) – é pioneiro para o português, ao tentar obter essa ontologia lexical semi-automaticamente a partir de um dicionário (neste caso também um dicionário restrito a uma variante, a de Portugal).

É importante contudo referir aqui que não o vemos como competidor, mas sim como mais uma contribuição para obter informação semântica de cobertura vasta para o português. Como descrito no resto deste artigo, consideramos que a situação ideal seria a de ter uma ontologia lexical pública para todo o português (embora entrando em conta e não desprezando as diferenças entre as variedades ou variantes da língua, veja-se por exemplo [Barreiro et al. 1996]); além disso, a forma de avaliação descrita na secção 4 é um bom início para uma ligação e para uma actualização de ambos os recursos envolvidos.

2. Contexto

Há basicamente três formas consagradas de construção de um recurso semântico de cobertura larga: (i) trabalho manual; (ii) através de processamento de corpos; e (iii) através do processamento de dicionários; apesar de novas ideias terem surgido nos últimos tempos, como através da análise de logs [Costa e Seco 2008] ou de jogos colaborativos. Embora a terceira forma seja a seguida no projecto PAPEL, usaremos recursos ou ideias obtidas pelos dois outros métodos, que mencionaremos brevemente aqui.

O PAPEL [Gonçalo Oliveira et al. 2008] foi construído a partir da análise automática das definições constantes numa versão electrónica do *Dicionário PRO da Língua Portuguesa* [dic 2005]. A utilização de dicionários em formato electrónico com vista à construção de recursos lexicais iniciou-se há cerca de quarenta anos, com estudos de [Calzolari et al. 1973] para o italiano e de [Amsler 1981] para o inglês. Desde aí se verificou que a estrutura dos dicionários e também a previsibilidade e simplicidade do vocabulário utilizado nas definições facilitariam a sua utilização para a extracção e organização de informação semântica. Apesar de vários trabalhos com este objectivo, já referidos por exemplo em [Gonçalo Oliveira et al. 2008], o MindNet [Richardson et al. 1998] terá sido a primeira base de dados lexical independente, criada de forma automática a partir de dicionários.

De longe a relação mais procurada (e achada) em abordagens usando corpos, é a relação de hiponímia. Veja-se, para o inglês, [Hearst 1992] e, para o português, [Freitas e Quental 2007]. Mas as relações causais ou finais, assim como a meronímia (parte de) e a localização geográfica também são e foram alvo de atenção pelas várias comunidades de extracção de informação, veja-se respectivamente [Girju e Moldovan 2002, Berland e Charniak 1999].

[Brank et al. 2005] apresentam quatro formas que têm sido utilizadas para avaliar ontologias de domínio: manualmente; comparação com um recurso dourado; realização de uma tarefa independente, definida para avaliar uma ontologia; ou comparação com um conjunto de dados sobre o mesmo domínio. Pela última forma, [Brewster et al. 2004] propõem a avaliação da adequação de uma ontologia a um dado texto através do número dos termos salientes num corpo de um dado domínio que também constam na ontologia. Estas ontologias, que cobrem apenas uma área específica, distinguem-se no entanto claramente das ontologias lexicais que tentam descrever o sistema conceptual de uma língua: repare-se que para obter os termos salientes num dado domínio, é necessário precisamente compará-lo com a linguagem geral e com o vocabulário de outros domínios, por isso obter os termos salientes na linguagem geral não faz sequer sentido.

Relativamente a ontologias lexicais construídas de forma automática, os autores do MindNet [Richardson et al. 1993] referem uma avaliação manual de 250 relações e

mais recentemente [Vanderwende et al. 2005] falam de uma avaliação da qualidade das relações que estará contudo ainda incompleta. [Nichols et al. 2005] utilizaram o WordNet e o GoiTaikei, um ”wordnet”para o japonês, como recursos dourados. Ao mesmo tempo que avaliaram a sua ontologia verificaram também que algumas relações se encontravam apenas num dos dois recursos dourados, o que pode indicar que estes estão incompletos.

3. Breve apresentação

De forma resumida, o processo de construção do PAPEL segue um ciclo de três passos até considerarmos ter chegado a um nível de desempenho suficiente, entrando depois no quarto e último passo.

1. **Criação de gramáticas semânticas**: na senda de [Alshawi 1989], foram criadas gramáticas para cada tipo de relação que se pretende extrair, por categoria gramatical (fornecida pelo dicionário). Na tabela 1 mostramos alguns dos padrões e as relações que pretendem descobrir e na figura 1 mostramos de que forma as relações que pretendemos extrair são descritas, de acordo com o grupo e especificando ainda a categoria dos argumentos e a sua relação inversa.

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

Tabela 1. Exemplos de padrões usados nas gramáticas.

```
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;
}
```

Figura 1. Exemplo da descrição do grupo de relações relativas à meronímia.

2. **A própria extracção**: usando um analisador sintáctico, é feita a análise superficial das definições, a partir da qual são automaticamente extraídas relações (descritas no passo anterior) entre palavras na definição e a palavra definida (ver figura 2).
3. **Inspecção dos resultados**: usando um sistema de regressão para identificar mais facilmente as diferenças entre resultados anteriores, procede-se à inspecção manual dos resultados obtidos, com a eventual volta ao primeiro passo para corrigir problemas detectados ou melhorar as gramáticas.
4. **Ajuste das relações**: aqui procuram-se corrigir (ou eliminar) de forma automática relações com argumentos inválidos.

O último passo é realizado em dois tempos. Inicialmente, todas as relações são transformadas no tipo directo[1]. Por exemplo, *manga* INCLUI *punho* é convertida para *punho* PARTE_DE *manga*, e *dor* RESULTADO_DE *distensão* é transformada em *distensão* CAUSADOR_DE *dor*.

[1]A escolha de um tipo directo e outro inverso foi arbitrariamente efectuada pelos criadores das gramáticas por um critério de naturalidade, e não de frequência, no dicionário ou em texto, e não tem qualquer consequência excepto a de facilitar a arrumação e depuração do recurso.

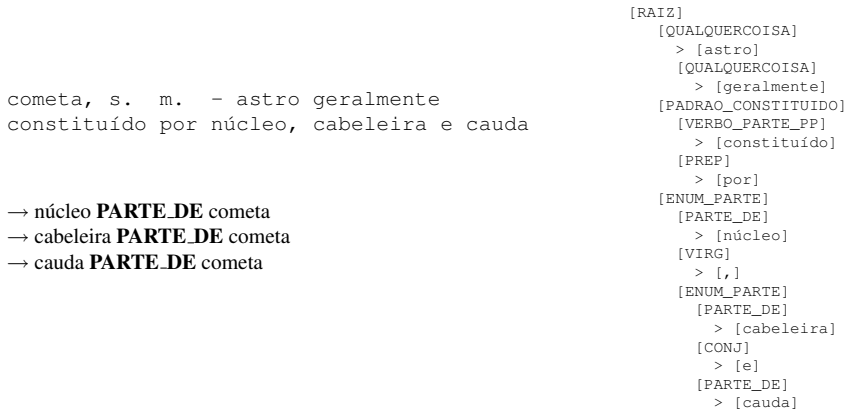

Figura 2. O resultado da análise da definição de *cometa*.

Visto que as gramáticas não fazem uma análise sintáctica das definições, não atribuindo por exemplo a classe gramatical, e que as definições do dicionário apenas incluem a classificação da vedeta, em alguns casos o processo de construção automática do PAPEL resulta em relações entre palavras de categorias erradas. É por isso preciso verificar, também de uma forma automática, esses casos, usando primeiro a própria lista de palavras/vedetas do dicionário e em seguida o analisador morfológico Jspell [Simões e Almeida 2002]. Se conseguirmos apurar que há uma desajuste nas categorias mas que pode ser corrigido através da escolha de outra relação pertencente ao mesmo grupo, substituímos, senão removemos esse triplo. Por exemplo, a relação *loucura* ACCAO_QUE_CAUSA *desvario* – que pressupõe um verbo como primeiro argumento – é transformada automaticamente em *loucura* CAUSADOR_DE *desvario*, visto que ambos os argumentos são substantivos. Durante este processo, os casos das palavras flexionadas são também substituídos pelos seus lemas, quando essa informação é dada pelo Jspell.

Após a sua construção, o PAPEL contém à volta de 200 000 relações, distribuídas de acordo com a tabela 2. A sinonímia e a hiperonímia são as relações mais frequentes, e ainda podem ser aumentadas, como discutiremos abaixo, de uma forma semelhante ao feito no ReRelEM [Freitas et al. 2009].

4. Avaliação

Aqui descrevemos uma primeira avaliação ao PAPEL, feita de duas formas diferentes. Dado que o Thesaurus Eletrônico para o Português do Brasil [Maziero et al. 2008] (Tep) pode ser levantado na rede, usámo-lo como recurso de referência para validar as relações de sinonímia, embora estejamos conscientes das várias diferenças entre as variantes. O Tep contém 19888 *synsets*, ou seja grupos de unidades lexicais com o mesmo sentido, correspondendo a 44678 unidades lexicais ao todo.

Para que a avaliação pudesse prosseguir sem enviesamento, começámos por retirar da comparação as entradas do Tep que não estivessem presentes no PAPEL assim como todas os casos de relações do PAPEL que contivessem argumentos ausentes do Tep. Ficámos assim apenas com 68% do nosso material, e com apenas 35% das possíveis

Grupo	Nome	arg1, arg2	Num.	Exemplo
Sinonímia	SINONIMO_DE	qq,=*arg1*	80432	(*flexível*, *moldável*)
Hiperonímia	HIPERONIMO_DE	sub,sub	63455	(*planta*, *salva*)
Meronímia	PARTE_DE	sub,sub	14453	(*cauda*, *cometa*)
	PARTE_DE_ALGO_COM_PROPRIEDADE	sub,adj	3715	(*tampa*, *coberto*)
	PROPRIEDADE_DE_ALGO_PARTE_DE	adj,sub	962	(*celular*, *célula*)
Causa	CAUSADOR_DE	sub,sub	1125	(*fricção*, *assadura*)
	CAUSADOR_DE_ALGO_COM_PROPRIEDADE	sub,adj	16	(*paixão*, *passional*)
	PROPRIEDADE_DE_ALGO_CAUSADOR_DE	adj, sub	515	(*reactivo*, *reacção*)
	ACCAO_QUE_CAUSA	v,sub	6424	(*limpar*, *purgação*)
	CAUSADOR_DA_ACCAO	sub,v	39	(*gases*, *fumigar*)
Produtor	PRODUTOR_DE	sub,sub	932	(*romãzeira*, *romã*)
	PRODUTOR_DE_ALGO_COM_PROPRIEDADE	sub,adj	31	(*sublimação*, *sublimado*)
	PROPRIEDADE_DE_ALGO_PRODUTOR_DE	adj,sub	348	(*fotógeno*, *luz*)
Fim	FINALIDADE_DE	sub,sub	2095	(*passagem_a_catedrático*, *agregação*)
	FINALIDADE_DE_ALGO_COM_PROPRIEDADE	sub,adj	23	(*enumeração*, *enumerativo*)
	ACCAO_FINALIDADE_DE	v,sub	5640	(*fazer_rir*, *comédia*)
	ACCAO_FINALIDADE_DE_ALGO_COM_PROP	v,adj	255	(*corrigir*, *correccional*)
	MANEIRA_POR_MEIO_DE	adv,sub	1433	(*timidamente*, *timidez*)
Lugar	LOCAL_ORIGEM_DE	sub,sub	768	(*Japão*, *japonês*)
Propriedade	PROPRIEDADE_DE_ALGO_REFERENTE_A	adj,sub	3700	(*dinâmico*, *movimento*)
	PROPRIEDADE_DO_QUE	adj,v	17028	(*diplomado*, *possuir_diploma*)

Tabela 2. As relações presentes no PAPEL.

405026 relações do Tep[2]. A comparação de ambos os conjuntos de relações produziu os seguintes resultados: 50% das nossas relações estavam presentes no Tep, e 39% das relações do Tep estavam presentes no PAPEL.

Embora estes valores possam ser surpreendentes, convém relembrar que as nossas relações tinham de ser encontradas directamente no dicionário, e não foram portanto ainda alvo de qualquer raciocínio. Em particular, a relação de transitividade parece ser óbvia: A SINONIMO_DE B $\wedge$ B SINONIMO_DE C $\rightarrow$ A SINONIMO_DE C. Após aplicação desta relação (uma vez só), obtivemos, dos 80432 sinónimos iniciais, 689073 sinónimos derivados. Claro está que, como as definições (e as nossas regras) não separam entre sentidos distintos de uma mesma palavra, esta expansão poderá levar a muitas relações infelizes, tal como *queda* SINONIMO_DE *ruína* $\wedge$ *queda* SINONIMO_DE *habilidade* $\rightarrow$ *ruína* SINONIMO_DE *habilidade*. Após esta expansão, e como esperado, o número de casos atestado no Tep caiu para 14%, contudo, 90% das relações no Tep puderam ser encontradas no PAPEL. Fica assim demonstrado que a combinação dos dois recursos permite não só melhorar ambos como separar o trigo do joio e mesmo alertar automaticamente para palavras com vários sentidos.

Em relação às outras relações, e na impossibilidade de comparar automaticamente com outros recursos para o português, tivemos de desenvolver uma metodologia diferente, inspirada nos vários trabalhos de extracção automática de relações semânticas em texto, ou de validação das mesmas em texto. Para os nossos testes usámos o CETEMPúblico [Rocha e Santos 2000], através da interface do projecto AC/DC [Costa et al. 2009], que nos permitiu além disso acesso às frequências dos lemas respectivos. O trabalho realizado tem de ser considerado preliminar, já que, devido a limitações de ocorrência de muitas das unidades lexicais nos corpos que usámos, não tivemos possibilidade de as testar. Com efeito, não só muitas das palavras no PAPEL eram demasiado raras ou especializadas, como cedo nos demos conta que

[2]Para conversão do Tep todos os elementos de um *synset* foram considerados como pertencendo a uma relação de sionónimia com todos os outros elementos do mesmo *synset*.

em texto jornalístico seria quase impossível encontrar num mesmo contexto (numa mesma frase) pares ou relações como *liquidar* ACCAO_QUE_CAUSA *liquidação*, *fósforo* PARTE_DE_ALGO_COM_PROPRIEDADE *fosforoso*, visto que são característicos de texto dicionarístico ou enciclopédico.

Restringimos assim o processo de validação, em primeiro lugar, apenas a relações entre substantivos, e, além disso, retirámos do teste as relações que envolvessem palavras cujos lemas estivessem ausentes do CETEMPúblico. Mesmo assim, e por questões de sobrecarga do serviço, para as duas relações mais populosas do PAPEL, hiperonímia e meronímia, ainda escolhemos uma amostra aleatória de relações a testar, correspondente a respectivamente 8% e 63% dos casos. Os resultados encontram-se na tabela 3.

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%

Tabela 3. Resultados da validação das relações excepto sinonímia.

Cerca de 20% destas relações parece serem validadas ou confirmadas pelo corpo, enquanto que a percentagem é menor para as outras relações. Estes resultados parecem-nos satisfatórios, tendo em conta que: o corpo é bastante pequeno; os padrões usados foram muito simples (em texto real há uma miríade de outras possibilidades de indicar uma relação); e os nossos valores não se encontram demasiado longe daqueles apresentados na literatura de confirmação. De qualquer maneira, e para mostrarmos que esta confirmação está longe de ser definitiva ou mesmo conclusiva, na tabela 4 apresentamos alguns exemplos, quer de confirmação certa quer de espúria (ou seja parecem confirmar mas não o fazem). Casos que não foram confirmados embora existam ambas as palavras no CETEMPúblico são por exemplo: *fruto* HIPERONIMO_DE *alperce*, *algoritmia* PARTE_DE *matemática*, *ausência* CAUSADOR_DE *saudade*, *aquecimento* FINALIDADE_DE *salamandra*.

Relação	Certa?	Justificação
língua HIPERONIMO_DE *italiano*	Sim	*As **línguas latinas, como o italiano** ou o português, tornam-se mais fáceis por causa das 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 vómitos e pode levar à desidratação e, consequentemente, 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.*
recrutamento FINALIDADE_DE *inspecção*	Sim	*Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às **inspecções para recrutamento**, revelou o ministro da Defesa.*
músico PARTE_DE *música*	Não	*... um espectáculo baseado na obra "Cantos de Maldoror", de Lautréamont, com **música composta pelo músico** inglês Steven Severin...*
fim FINALIDADE_DE *sempre*	Não	*Sicilia aponta **sempre para o fim** do dia, para o fim da luz.*

Tabela 4. Exemplos de validação através do CETEMPúblico.

5. Comentários finais

Apresentamos neste artigo um novo recurso lexical para o português, que poderá ser levantado integralmente no endereço acima citado. O PAPEL não pretende ser um recurso

final, mas sim um ponto de partida importante para futuros projectos, que o poderão enriquecer recorrendo a outras fontes de informação.

A sua primeira avaliação, apesar de bastante preliminar, pode também ser interessante como exemplo de avaliação de outros recursos. Pretendemos no futuro continuar a melhorar o PAPEL através da obtenção de novos dados na rede assim como aperfeiçoar a validação dos já presentes.

Agradecimentos

Agradecemos ao grupo de R&D da Porto Editora por nos ter disponibilizado o dicionário. O projecto PAPEL é desenvolvido no âmbito da Linguateca, co-financiada pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contracto POSC/339/1.3/C/NAC, pela UMIC e pela FCCN. Hugo Gonçalo Oliveira é financiado pela FCT, bolsa SFRH/BD/44955/2008.

Referências

(2005). *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto.

H. Alshawi (1989). Analysing the dictionary definitions. *Computational lexicography for natural language processing*, páginas 153–169.

Robert A. Amsler (1981). A taxonomy for english nouns and verbs. Em *Proc. the 19th annual meeting on Association for Computational Linguistics*, páginas 133–138, Morristown, NJ, USA. Association for Computational Linguistics.

Anabela Barreiro, Luzia Helena Wittmann e Maria de Jesus Pereira (1996). Lexical differences between European and Brazilian Portugueses. *INESC Journal of Research and Development*, 5(2).

Matthew Berland e Eugene Charniak (1999). Finding parts in very large corpora. Em *Proc. 37th Annual Meeting of the ACL on Computational Linguistics*, páginas 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

Janez Brank, Marko Grobelnik e Dunja Mladenic' (2005). A survey of ontology evaluation techniques. Em *Proc. Conference on Data Mining and Data Warehouses (SiKDD)*.

Christopher Brewster, Harith Alani, Srinandan Dasmahapatra e Yorick Wilks (2004). Data-driven ontology evaluation. Em *Proc. the Language Resources and Evaluation Conference (LREC)*, páginas 164–168, Lisbon, Portugal. European Language Resources Association.

Nicoletta Calzolari, Laura Pecchia e Antonio Zampolli (1973). Working on the italian machine dictionary: a semantic approach. Em *Proc. 5th conference on Computational linguistics*, páginas 49–52, Morristown, NJ, USA. Association for Computational Linguistics.

Luís Costa, Diana Santos e Paulo Rocha (2009). Estudando o português tal como é usado: o serviço AC/DC. Neste volume.

Rui P. Costa e Nuno Seco (2008). Hyponymy extraction and web search behavior analysis based on query reformulation. Em *Proc. 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA)*, LNAI, páginas 332–341. Springer Verlag.

Bento C. Dias da Silva, Mirna Oliveira e Helio Moraes (2002). Groundwork for the Development of the Brazilian Portuguese Wordnet. Em Nuno Mamede e Elisabete Ranchhod, editores, *Proc. Advances in Natural Language Processing: 3rd International Conference*, LNAI, páginas 189–196. Springer Verlag.

Christiane Fellbaum, editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Cláudia Freitas e Violeta Quental (2007). Subsídios para a elaboração automática de taxonomias. Em *XXVII Congresso da SBC - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, páginas 1585–1594.

Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira e Paula Carvalho (2009). Detection of relations between named entities: report of a shared task. Em *Proc. NAACL-HLT Workshop, Semantic Evaluations: Recent Achievements and Future Directions*.

Roxana Girju e Dan Moldovan (2002). Text mining for causal relations. Em Susan M. Haller e Gene Simmons, editores, *Proc. 15th Intl. Florida Artificial Intelligence Research Society Conference (FLAIRS)*, páginas 360–364.

Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes e Nuno Seco (2008). PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira et. al, editor, *Proc. of Computational Processing of the Portuguese Language, 8th Intl. Conf. (PROPOR)*, volume 5190 de *LNAI*, páginas 31–40. Springer Verlag.

Marti A. Hearst (1992). Automatic acquisition of hyponyms from large text corpora. Em *Proc. the 14th conference on Computational linguistics*, páginas 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

Palmira Marrafa (2002). Portuguese wordnet: general architecture and internal semantic relations. *Documentação de Estudos em Linguística Teórica e Aplicada (DELTA)*, 18:131–146.

Erick G. Maziero, Thiago A. S. Pardo, Ariani Di Felippo e Bento C. Dias-da-Silva (2008). A base de dados lexical e a interface web do tep 2.0 - thesaurus eletrônico para o português do brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, páginas 390–392.

Eric Nichols, Francis Bond e Dan Flickinger (2005). Robust ontology acquisition from machine-readable dictionaries. Em Leslie Pack Kaelbling e Alessandro Saffiotti, editores, *IJCAI*, páginas 1111–1116. Professional Book Center.

Stephen D. Richardson, William B. Dolan e Lucy Vanderwende (1998). Mindnet: acquiring and structuring semantic information from text. Em *Proc. 17th Intl. Conference on Computational linguistics*, páginas 1098–1102, Morristown, NJ, USA. Association for Computational Linguistics.

Stephen D. Richardson, Lucy Vanderwende e William Dolan (1993). Combining dictionary-based and example-based methods for natural language analysis. Em *Proc. 5th Intl. Conference on Theoretical and Methodological Issues in Machine Translation*, páginas 69–79, Kyoto, Japan.

Paulo Alexandre Rocha e Diana Santos (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR)*, páginas 131–140, São Paulo. ICMC/USP.

Alberto M. Simões e J.J. Almeida (2002). Jspell.pm – um módulo de análise morfológica para uso em processamento de linguagem natural. Em *Actas do XVII Encontro da Associação Portuguesa de Linguística*, páginas 485–495, Lisboa.

Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki e Arul Menezes (2005). Mindnet: An automatically-created lexical resource. Em *Proc. of HLT/EMNLP on Interactive Demonstrations*. The Association for Computational Linguistics.