ESTUDOS DE LINGUÍSTICA

VOLUME I

ANA R. LUÍS
COORD.

IMPRENSA DA UNIVERSIDADE DE COIMBRA
COIMBRA UNIVERSITY PRESS

Hugo Gonçalo Oliveira e Paulo Gomes
*Centro de Informática e Sistemas da Universidade de **Coimbra** (**CISUC**)*

Onto.PT: Construção Automática de uma Ontologia Lexical para o Português

1. Introdução

As bases de conhecimento lexical de grande cobertura têm um papel cada vez mais importante no desenvolvimento de ferramentas computacionais que necessitam de interpretar informação transmitida através de linguagem natural. Para o inglês, a WordNet de Princeton, WN.Pr (Fellbaum 1998), é amplamente utilizada na realização de tarefas de processamento de linguagem natural (PLN), desde a determinação de semelhanças (Agirre *et al.* 2009), passando pela desambiguação do sentido das palavras (Resnik 1995) ou pela resposta automática a perguntas (Pasca e Harabagiu 2001).

É indiscutível que a existência de uma *wordnet* potencia o desenvolvimento de ferramentas de PLN para a sua língua, daí a expansão desta forma de recurso para várias línguas. Ainda assim, há que referir que a existência de uma *wordnet* não implica o fim da investigação na extracção automática de conhecimento léxico-semântico. Mais uma vez, para o inglês, há trabalhos que têm, por exemplo, o objectivo de ampliar a WN.Pr (veja-se (Hearst 1998)).

Neste artigo é apresentado o Onto.PT (Gonçalo Oliveira e Gomes 2010), um projecto que tem o objectivo de criar uma alternativa pública dentro do paradigma das ontologias lexicais para o português. Esta ontologia será construída de forma automática e estruturada de forma semelhante a uma *wordnet*.

Muito resumidamente, numa primeira fase de construção, são extraídas instâncias de relações semânticas, e numa segunda procuram-se aglomerados

de palavras relacionadas por sinonímia de maneira a formar conceitos. Por fim, procuram-se os conceitos mais adequados a cada argumento, uma palavra, das demais relações extraídas. Apesar de, na primeira versão do Onto.PT, terem apenas sido explorados dicionários e *thesauri*, o processo de construção é suficientemente flexível para, no futuro, poder incorporar conhecimento de outros recursos, tais como a Wikipédia.

Para além de apresentarmos, brevemente, a metodologia seguida na construção do Onto.PT, este artigo foca-se em vários exemplos, não só para complementar a explicação de cada uma das fases de construção, mas também para ilustrar alguns dos resultados mais interessantes a que, até ao momento, chegamos.

Além de uma nova rede lexical e de um novo *thesaurus* electrónico público para o português, que acreditamos serem os maiores do seu tipo, as principais contribuições deste projecto são um novo recurso lexical para o português, e uma metodologia flexível que poderá ser adaptada à construção de recursos semelhantes para outras línguas, e que proporciona ao Onto.PT um grande potencial de crescimento.

Iniciamos este artigo com o enquadramento deste trabalho, onde referimos alguns trabalhos relacionados e recursos semelhantes. Depois de explicitarmos a abordagem adoptada na construção do Onto.PT, apresentamos alguns resultados obtidos, dando exemplos de relações extraídas e *synsets* descobertos.

2. Enquadramento

A necessidade de realizar tarefas de PLN onde, para além de reconhecer as palavras e as suas interacções, é crucial interpretar o significado do texto, levou à criação de recursos léxico-semânticos de ampla cobertura, tais como as ontologias lexicais, recursos que têm ao mesmo tempo propriedades de um léxico e de uma ontologia (Hirst 2004, Prévot *et al.* 2010).

Do ponto de vista da engenharia informática, uma ontologia é um sistema formal baseado em conceitos, regra geral de um domínio, e relações entre eles. Por outro lado, um léxico é um recurso linguístico que contém

conhecimento sobre uma língua e inclui, por isso, itens lexicais, como as palavras. No entanto, como, na realidade, as relações semânticas não se estabelecem entre palavras, mas entre significados (sentidos de palavras), e como os significados são inerentemente conceptuais, as bases de conhecimento lexical podem ser vistas como ontologias lexicais, onde conceitos da linguagem natural se encontram descritos através de palavras.

Como a construção de uma ontologia é uma tarefa tipicamente complexa que requer muito tempo, vários autores desde cedo estudaram formas de automatizar este processo. Tendo em conta a sua estrutura, os dicionários electrónicos foram dos primeiros alvos a ser explorados na construção automática de ontologias lexicais. Após analisar as regularidades presentes nas definições, Amsler (1981) chegou à conclusão que não seria complicado desenvolver procedimentos automáticos que, tirando partido de padrões lexicais ou léxico-sintácticos empregues, permitissem a extracção de relações de hiperonímia (Chodorow *et al.* 1985) ou mesmo a construção de bases de conhecimento lexical, como a MindNet (Richardson *et al.*, 1998), que inclui um vasto conjunto de relações semânticas.

Apesar de tudo, foram apontados problemas à utilização de um dicionário para esta tarefa (Ide e Veronis 1995). Para além de nem sempre estarem disponíveis para fins de investigação, verificou-se que dicionários diferentes apresentam normalmente a informação de forma diferente (por exemplo, não cobrem exactamente da mesma forma as mesmas partes do léxico) e remover, por vezes, até se complementam. Assim, a utilização de vários dicionários é vista como uma forma de ultrapassar alguns dos problemas identificados.

Com o estabelecimento da WN.Pr como uma base de conhecimento lexical amplamente utilizada no inglês, a investigação na extracção automática deste tipo de conhecimento a partir de dicionários diminuiu consideravelmente. Ainda assim, os dicionários continuam a ser utilizados na construção de ontologias (Nichols *et al.* 2005), onde se inclui o projecto PAPEL (Gonçalo Oliveira *et al.* 2010), que constitui uma rede lexical para o português.

A WN.Pr é um recurso criado manualmente por especialistas e baseado em *synsets* – grupos de palavras sinónimas que traduzem conceitos da linguagem natural. Cada *synset* tem uma definição, semelhante às entradas de um dicionário, sendo possível estabelecer várias relações semânticas

(e.g. hiperonímia, parte-de) entre *synsets*. Além de ser uma alternativa interessante à representação computacional de palavras e sentidos, a WN.Pr tornou-se num recurso amplamente utilizado devido ao seu carácter público. No entanto, sendo uma base de conhecimento lexical, não contempla conhecimento específico de determinados domínios, o que levou a que investigadores a tenham enriquecido com conhecimento extraído de outros recursos, como corpora (veja-se (Hearst 1998), (Lin e Pantel 2002)) ou enciclopédias, onde se inclui a Wikipédia (ver (Medelyan *et al.* 2009)).

O sucesso da WN.Pr levou a que fosse adaptada a outras línguas, incluindo as que se inserem nos projectos EuroWordNet (Vossen 1997) e MultiWordNet (Pianta *et al.* 2002). No entanto, a criação de uma *wordnet* requer muito tempo e envolve um grande volume de trabalho manual. Isto levou a que autores propusessem a tradução, de forma automática, da WN.Pr (Melo e Weikum 2008) para outras línguas. No entanto, e apesar desta abordagem provavelmente se adequar a algumas tarefas, línguas diferentes representam diferentes realidades socio-culturais: não cobrem exactamente a mesma parte do léxico e, por vezes, conceitos diferentes são lexicalizados de forma diferente (Hirst 2004). Assim, acreditamos que a construção da *wordnet* para uma língua deve ser desenvolvida de raiz para essa língua.

Para a língua portuguesa existem actualmente duas *wordnets* proprietárias, a WordNet.PT[99] (Marrafa 2002) e a MultiWordNet.PT (MWN.PT)[100]. Ambas (ou partes de ambas) estão disponíveis para pesquisa através da rede. Além disso, a MWN.PT, que se encontra alinhada com a WN.Pr, pode ser adquirida. Ambas sofrem também pelo seu método de construção manual, que os limita em termos de crescimento e cobertura (Santos *et al.* 2010). No contexto das ontologias lexicais, existem mais alguns recursos para a o português que devem ser mencionados:

[99] http://www.clul.ul.pt/clg/eng/wordnetpt/index.html

[100] http://mwnpt.di.fc.ul.pt/

- Dois *thesauri*: TeP[101] (Maziero *et al.* 2008), para a variante brasileira, que acreditamos ser a base de *synsets* para a *wordnet* brasileira (WordNet. Br (Dias-da-Silva *et al.* 2002)); e o OpenThesaurus.PT[102], um *thesaurus* colaborativo utilizado no processador de texto do OpenOffice[103] para sugestão de sinónimos;
- Wikcionário[104], um dicionário colaborativo onde, além de definições, é possível adicionar informação acerca de relações semânticas para cada entrada. No entanto, a versão portuguesa deste recurso é ainda pequena e limitada;
- PAPEL[105] (Gonçalo Oliveira *et al.* 2010), uma rede lexical pública que contém instâncias de vários tipos de relações semânticas, extraídas de forma automática a partir da edição de 2005 do Dicionário da Língua Portuguesa (DLP 2005). As principais diferenças entre o PAPEL e uma *wordnet* são a construção automática do PAPEL e a sua estrutura baseada em palavras, ao contrário de *synsets*.

Mais informação sobre alguns destes recursos pode ser encontrada em (Santos *et al.* 2010). Uma comparação focada nos verbos encontra-se em (Teixeira *et al.* 2010). Ambos os trabalhos de comparação verificaram que, apesar de se tratarem assumidamente de recursos léxico-semânticos de grande cobertura, o seu conteúdo se complementa. Logo, a utilização não de um, mas de vários destes recursos, seria positiva na construção de um recurso único, com maior cobertura. Este problema parece ser comum a outras línguas, tais como a inglesa, onde, como referido anteriormente, há vários trabalhos com o objectivo de enriquecer a WN.Pr.

[101] http://www.nilc.icmc.usp.br/tep2/

[102] http://openthesaurus.caixamagica.pt/

[103] http://www.openoffice.org/

[104] http://pt.wiktionary.org/

[105] http://www.linguateca.pt/PAPEL/

3. Abordagem

O Onto.PT é uma ontologia lexical para o português, estruturada de forma semelhante a uma *wordnet*, que pretende ser uma nova alternativa no campo dos recursos lexicais para a nossa língua. É criada de forma automática a partir de recursos textuais, seguindo uma abordagem flexível

que permite a integração de conhecimento presente em vários recursos, tais como os apresentados anteriormente. A abordagem para a construção do Onto.PT consiste num procedimento automático com três fases, que se descrevem de seguida.

3.1. Extracção de relações

Nesta fase é necessário construir, manualmente, um conjunto de gramáticas com padrões textuais indicadores de um conjunto pré-definido de relações semânticas. O texto é depois processado por um analisador sintáctico que utiliza as gramáticas na extracção automática de instâncias de relações, representadas através de triplos *t=(a,R,b)*, onde *a* e *b* são palavras e *R* é o nome de uma relação estabelecida entre um sentido de *a* e um sentido de *b*.

Para a primeira versão do Onto.PT, apenas foram processadas definições de dicionários. O procedimento de extracção utilizado é inspirado na construção do PAPEL (ver (Gonçalo Oliveira *et al.* 2010)). Sendo assim, a partir de definições como:

> **candeia** s.f. utensílio doméstico rústico usado para iluminação, com pavio abastecido a óleo
> **espiga** s.f. parte das gramíneas que contém os grãos
> **inquietar** v.t. causar ansiedade
> **severo** adj. grave , crítico

É possível tirar partido dos padrões textuais sublinhados para extrair, por exemplo, os seguintes triplos:

utensílio HIPERONIMO_DE *candeia*
iluminação FINALIDADE_DE *candeia*
espiga PARTE_DE *gramínea*
grão PARTE_DE *espiga*
inquietar CAUSADOR_DE *ansiedade*
grave SINONIMO_DE *severo*
crítico SINONIMO_DE *severo*

3.2. Descoberta de synsets

Ao olhar apenas para as relações de sinonímia, verifica-se que estas tendem a formar aglomerados (vulgo *clusters*) de palavras. Nesta fase, aglomerados são identificados de forma automática e aproximados a *synsets*, de acordo com um procedimento semelhante ao apresentado em (Gonçalo Oliveira e Gomes 2011a). O resultado é um *thesaurus*, onde cada conceito é representado por um conjunto de palavras sinónimas, à imagem de uma *wordnet*.

As figuras 1 e 2 apresentam dois exemplos de redes de palavras, formadas por relações de sinonímia extraídas automaticamente. Cada ligação entre duas palavras indica que foi extraída uma relação de sinonímia entre ambas, e fundos com diferentes tonalidades de cinzento identificam aglomerados. Na figura 1 existe uma ambiguidade relacionada com o substantivo 'histórico', que tanto se pode referir a uma cronologia, como a alguém que estuda história, um historiador. A ambiguidade acaba por ser desfeita e, devido aos parâmetros utilizados, 'histórico' fica apenas com o significado de cronologia.

Por outro lado, a figura 2 apresenta um exemplo mais complexo, numa rede maior, onde existem várias ambiguidades. No fim, é possível identificar três conceitos que, apesar de próximos, são diferentes: (A) alguém que muda de local em busca de melhor sorte; (B) alguém que foge do seu país por estar a ser perseguido; e ainda (C) alguém que foi expulso do seu país. Também na figura 2, verifica-se que algumas palavras mantêm alguma ambiguidade e pertencem a dois ou três aglomerados. Isto vai ao encontro

da realidade das linguagens naturais, onde a mesma palavra pode efectivamente ter mais de um significado.

3.3. Integração das relações

Por fim, procuram-se associar os argumentos das restantes relações extraídas, aos synsets descobertos anteriormente, de acordo com um dos métodos apresentados em (Gonçalo Oliveira e Gomes 2011b).

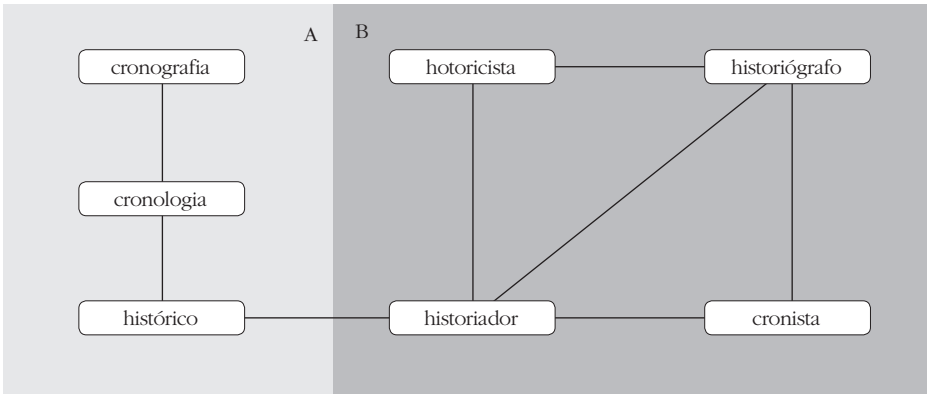

Figura 1: Rede de sinonímia em torno da palavra "histórico" e aglomerados identificados

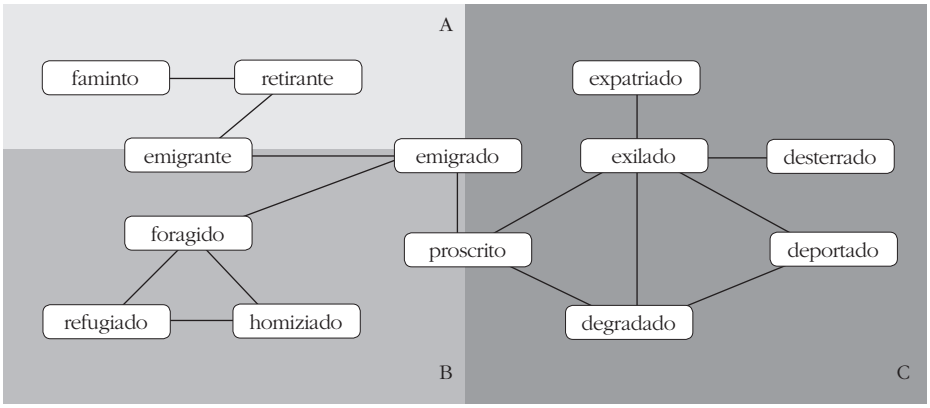

Figura 2: Rede de sinonímia em torno da palavra "emigrado" e aglomerados identificados

Por outras palavras, os triplos que relacionam palavras passam a relacionar conceitos, ainda que descritos por palavras. Um dos métodos que utilizamos na realização desta tarefa procura o par de *synsets* mais semelhante, dentro de todos os pares possíveis de *synsets* (S_a, S_b), em que S_a contém o termo *a*, e S_b contém o termo *b*. A semelhança é calculada com base nas vizinhanças dos termos dos *synsets* numa rede formada por todos os triplos extraídos. A tabela 1 apresenta alguns exemplos de relações entre termos incluídos no PAPEL e a sua correspondência após a integração num *thesaurus*, o TeP.

Triplos de termos	Triplos de synsets
aparelho **Hiperónimo-de** *televisor*	*apresto,utensílio,petrechos,instrumento,apetrechos,aparelho* **Hiperónimo-de** *tv,televisão,televisor,têvê*
extensão **Hiperónimo-de** *território*	*superfície,dimensão,extensão;espaço,área* **Hiperónimo-de** *território,área*
ângulo **Parte-de** *triângulo*	*ângulo,face,lado* **Parte-de** *triângulo,trilateral*
técnica **Membro-de** *marketing*	*arte,técnica* **Membro-de** *marketing*
edição **Finalidade-de** *programa*	*edição,lançamento* **Finalidade-de** *programa,aplicativo*

Tabela 1. Exemplos da integração de relações no thesaurus.

4. Resultados obtidos

Nesta secção são apresentados os resultados mais recentemente obtidos na criação do Onto.PT. Até ao momento, foram explorados três dicionários: o Dicionário da Língua Portuguesa (DLP, através do PAPEL 2.0); o Dicionário Aberto (DA, (Simões e Farinha 2010)), um dicionário público do português, com data de 1913, recentemente convertido para um formato digital; e também a versão portuguesa do Wikcionário. Foram ainda utilizados dois

thesauri, mais propriamente o TeP e o OpenThesaurus.PT. Na construção aqui referida, os *synsets* foram transformados em pares de sinonímia e, antes da descoberta de *synsets*, acrescentados às relações de sinonímia extraídas dos dicionários[106].

Iniciamos esta secção com uma análise ao vocabulário utilizado nas definições dos dicionários, crucial para a criação das gramáticas utilizadas na extracção de relações semânticas. Depois, apresentamos as relações extraídas, dados sobre os *synsets* descobertos acompanhados de exemplos e, por fim, dados sobre a versão actual do Onto.PT.

4.1. Análise de vocabulário

Antes de construir as gramáticas utilizadas na extracção de relações semânticas, é necessário analisar o vocabulário utilizado nas definições de dicionários para, assim, identificar mais facilmente os padrões que podem e devem ser utilizados na tarefa anteriormente referida. A tabela 2 mostra alguns dos padrões mais frequentes nas definições dos dicionários explorados, a sua quantidade em cada dicionário e ainda a relação semântica que normalmente transmitem.

Além de padrões como os da tabela 2, as gramáticas tiram ainda partido de outras duas construções, nomeadamente:

- Definições com apenas uma palavra, ou uma enumeração de palavras, são utilizadas na extracção de relações de sinonímia entre as palavras da definição e a palavra definida.
- Uma grande parte das definições de substantivos iniciam-se por um hiperónimo da palavra definida, eventualmente modificado por um adjectivo, a menos que essa palavra seja uma "cabeça vazia", ou seja, uma palavra sem conteúdo como, por exemplo, 'tipo' ou 'parte' (ver (Chodorow *et al.* 1985)).

[106] Além do procedimento aqui descrito, temos experimentado outras formas de integrar synsets existentes no Onto.PT. (Gonçalo Oliveira e Gomes 2011c).

Padrão	Categoria	Frequência			Relação semântica
		DLP	DA	Wikcionário	
o mesmo que	Substantivo	0	10.627	960	Sinonímia
a[c]to ou efeito de	Substantivo	3.851	2.501	553	Causa
aquele que	Substantivo	1.148	3.357	476	Hiperonímia
conjunto de	Substantivo	1.004	316	293	Membro
espécie de	Substantivo	798	2.846	202	Hiperonímia
género de	Substantivo	29	4.148	47	Hiperonímia
variedade de	Substantivo	455	621	42	Hiperonímia
[a] parte do/da	Substantivo	445	433	96	Parte
natural ou habitante de/do/da	Substantivo	528	0	0	Local/Origem
habitante ou natural de	Substantivo	6	0	211	Local/Origem
instrumento[,] para	Substantivo	94	284	17	Finalidade
.. produzid[o/a] por/pelo/pela	Substantivo	155	146	31	Produtor
o mesmo que	Verbo	0	166	84	Sinonímia
Fazer	Verbo	1.680	1.294	328	Causa
Tornar	Verbo	1.359	1.672	277	Causa
Ter	Verbo	467	519	154	Propriedade
o mesmo que	Adjectivo	0	2.685	179	Sinonímia
relativo a/á/ao	Adjectivo	1.236	5.554	946	Propriedade
que se	Adjectivo	1,602	1.599	443	Propriedade
que tem	Adjectivo	2.698	4,291	437	Parte/Propriedade
diz-se de	Adjectivo	2.066	738	272	Propriedade
relativo ou pertencente	Adjectivo	1.647	9	50	Membro/ Propriedade
habitante ou natural de	Adjectivo	0	0	143	Local/Origem
de modo	Advérbio	398	2,261	89	Maneira
de maneira	Advérbio	49	9	28	Maneira
de forma	Advérbio	30	3	18	Maneira

Tabela 2: Padrões frequentes e produtivos em definições de dicionários.

Ao realizarmos esta análise, verificamos que, com pequenas adaptações, as gramáticas utilizadas na construção do PAPEL 2.0[107] poderiam ser reutilizadas, já que grande parte dos padrões utilizados no DLP são também empregues nos outros dois dicionários.

[107] Disponíveis livremente a partir de http://www.linguateca.pt/PAPEL/

4.2. Relações extraídas

Ao reutilizarmos as gramáticas do PAPEL, no Onto.PT acabamos por extrair o mesmo tipo de relações presentes no primeiro recurso. Na tabela 3, apresentamos os vários tipos de relação extraídos, que separamos ainda em sub-relações, tendo em conta a categoria gramatical dos argumentos. Ainda na tabela 3, apresentamos as quantidades destas sub-relações extraídas e ainda exemplos para cada uma delas.

Relação	Args.	Quantidade				Exemplos
		PAPEL	DA	Wikc.	Total	
Sinonímia	n,n	37.452	24.327	13.569	64.131	*alegria,satisfação*
	v,v	21.465	11.031	4.076	30.380	*esticar,estender*
	adj,adj	19.073	10.114	6.640	29.676	*racional,filosófico*
	adv,adv	1.171	1.187	289	2.292	*imediatamente,já*
Hiperonímia	n,n	62.591	54.059	17.845	108.900	*sentimento,afecto*
Parte-de	n,n	2.805	1.280	718	4.485	*núcleo,átomo*
	n,adj	3.721	3.165	558	6.058	*vício,vicioso*
	adj,n	17	45	26	87	*sujeito,oração*
Membro-de	n,n	5.929	1.259	1.153	7.889	*aluno,escola*
	n,adj	34	25	23	77	*coisa,coletivo*
	adj,n	883	84	121	1.071	*rural,campo*
Causa-de	n,n	1.013	172	306	1.440	*vírus,doença*
	n,adj	17	7	4	23	*paixão,passional*
	adj,n	498	142	163	735	*horrível,horror*
	n,v	39	17	5	59	*fogo,fundir*
	v,n	6.399	5.487	1.139	8.812	*mover,movimento*
Produtor-de	n,n	898	571	310	1.652	*oliveira,azeitona*
	n,adj	35	25	8	65	*fermentação,fermentado*
	adj,n	359	230	37	520	*fonador,som*
Finalidade-de	n,n	2.886	2.302	1.349	6.194	*sustentação,mastro*
	n,adj	63	48	9	111	*habitação,habitável*
	v,n	5.192	2.186	1.479	8.314	*calcular,cálculo*
	v,adj	260	197	9	403	*comprimir,compressivo*
Local-de	n,n	849	391	728	1.858	*Galiza,galego*
Maneira-de	adv,n	1.113	1.535	156	2.482	*ociosamente,indolência*
	adv,adj	N/A	1.594	112	1.664	*virtualmente,virtual*
Maneira	adv,n	117	148	24	260	*prontamente,demora*
sem	adv,v	11	4	5	19	*seguido,parar*
Propriedade-de	adj,n	6.518	3.460	1.636	10.226	*daltónico,daltonismo*
	adj,v	17.543	9.701	3.239	27.916	*geoquímico,geoquímica*

Tabela 3: Relações semânticas extraídas

4.3. Synsets descobertos

Ao observar as redes estabelecidas não só pelas relações de sinonímia acima descritas, mas também por pares de sinonímia obtidos a partir dos *thesauri* TeP e OpenThesaurus.PT, verificou-se que estas redes eram constituídas por uma grande sub-rede de palavras ligadas e por várias pequenas sub-redes. A tabela 4 apresenta alguns números relativamente ao tamanho das sub-redes para cada categoria gramatical, nomeadamente, o número total de palavras da categoria, o total de sub-redes, o tamanho da maior e segunda maior sub-redes e ainda o número de sub-redes com apenas duas palavras.

Olhando para estes números, verificamos que, se não existisse ambiguidade, todas as palavras de cada sub-rede seriam sinónimas entre si, porque é possível encontrar um caminho de relações de sinonímia entre qualquer par de palavras na mesma sub-rede. No entanto, é necessário descobrir *synsets* dentro das sub-redes e assim desfazer ou tratar algumas ambiguidades.

A descoberta automática de *synsets* resulta num *thesaurus* cujas propriedades se apresentam na tabela 5, para cada categoria gramatical. As propriedades incluem, para as palavras, o número total de palavras, palavras com mais de um sentido (Ambigs.), o número médio de sentidos por palavra (med(sents)) e o número de sentidos da palavra mais ambígua (max(sents)). No que diz respeito aos *synsets*, é apresentado o número total, o número médio de palavras por *synset* (med(pals)), o número de *synsets* com duas palavras (*pals=2*), com mais de 25 palavras (*pals>25*) e ainda o número de palavras no maior *synset* (max(pals)).

Categoria	Palavras	Sub-redes			
		Total	Tamanho maior	Tamanho 2ª maior	Apenas 2 palavras
n	45.214	5.597	30.922	48	4.016
v	11.975	345	11.186	6	277
adj	21.840	2.928	14.587	28	2.155
adv	2.504	215	1.894	10	135

Tabela 4: Dados sobre as sub-redes de sinonímia extraídas.

POS	Palavras				Synsets				
	Total	Ambigs	med (sents)	max (sents)	Total	med (pals)	pals=2	pals>25	max (pals)
n	45.213	17.717	1,75	8	14.057	5,63	4.178	48	46
v	11,904	6.612	1,96	8	3.549	6,58	571	46	55
adj	21.837	8.695	1,73	8	6.550	5,78	2.303	61	48
adv	2.504	876	1,61	7	651	6,19	135	1	27

Tabela 5: Propriedades do thesaurus

De forma a verificar como o *thesaurus* gerado automaticamente lida com a polissemia, procuramos por diferentes *synsets* de palavras claramente polissémicas, que tentamos separar pelos diferentes sentidos que essas palavras transmitem. Embora nem todos os sentidos estivessem representados para todas as palavras e, por vezes, houvesse *synsets* que misturavam mais do que um sentido, há bons exemplos, como os representados na tabela 6. Aí, temos a palavra: 'etiqueta', que pode ter o sentido de uma legenda/rótulo ou de um protocolo que deve ser seguido, e 'pastel', com dois dos seus possíveis sentidos figurados, nomeadamente uma pessoa indolente e também dinheiro.

Palavra	Sentido	Synsets
etiqueta	legenda	*etiqueta, signal, letreiro, rótulo, dístico, tabuleta, matrícula, tãtuleiro, cartel, inscrição, epitáfio, legenda*
	protocolo	*etiqueta, praxe, práxis, protocolo, cerimónia, pragmática, cerimônia, formalidade, requifife, costumeira, usança, premática*
pastel	indolente	*pastel, zorrão, marralhão, indolente, mandriona, boleima*
	dinheiro	*pastel, guines, moeda, jimbo, bagalhoça, cobre, boro, dieiro, baguines, parrolo, marcaureles, cacau, pataco, matambira, massaroca, gimbo, metal, cunques, bagalho, níquel, fanfa, bilhestres, pecúnia, jan-da-cruz, cum-quibus, mussuruco, pilim, dinheiro, zerzulho, numo, chelpa, calique, teca, pecunia, patacaria, carcanhol, pecuniária*

Tabela 6: Synsets de palavras polissémicas.

O *synset* relativo a dinheiro é um dos maiores do *thesaurus* obtido e, de acordo com os parâmetros utilizados, poderá ainda ser maior, como é o caso do *synset* equivalente apresentado em (Gonçalo Oliveira e Gomes

2011a). Neste *synset* encontramos várias formas de nos referirmos a dinheiro, como por exemplo: formas coloquiais (e.g. pastel, carcanhol, pilim), formas populares (e.g. massaroca, cacau), formas da variante moçambicana do português (e.g. mussuruco, matambira), formas figurativas (metal), gírias (e.g. baguines, gimbo) ou formas mais antigas (dieiro).

Além deste *synset*, apresentamos de seguida, a título de curiosidade, os maiores *synsets* obtidos para substantivos e para verbos, juntamente com o significado que estes transmitem. Observa-se que os maiores *synsets* de verbos têm mais palavras que os *synsets* de substantivos.

Substantivos:

confusão (46 palavras): *baixaria, furdúncio, matalotagem, fuzuê, confusão, complicação, feijoada,mistura, encrenca, remexida, amalgamação, salada_russa, saricoté, brigas, estrilho, escangalho, abstrusão, trapalhada, siricutico, vira-teimão, cambulha, cu-de-boi, valverde, kanvuanza, javardice, embrolho, canvanza, indistincção, estricote, caravançarai, imbróglio, desmanbo, ensalsada, enredia, vuvu, mexedura, cancaburra, ula, fula-fula, timaca, misturada, pastelada, lelê, abstrusidade, assarapantamento, encrequilha*

embriaguez (38 palavras): *pisorga, porre, berzundela, zurca, zuca, trapisonda, torta, piela, bebedeira, perua, perunca, bicancra, gateira, taçada, carraspana, samatra, pizorga, tachada, ganso, tortelia, caroça, turca, pifão, borracheira, carrega, pifo, zerenamora, zola, marta, parrascana, gardinhola, tropecina, bezana, ema, cardina, lavanco, tiorga, bêbeda*

prostituta (38 palavras): *pécora, galinha-polaca, meretriz, faniqueira, barregã, mulher-dama, mulher-do-fandango, mulher-de-má-nota, dadeira, mulher-perdida, zoupeira, arruadeira, hervoeira, messalina, mundana, prostituta, michê, tronga, desproveito, prostituída, culatrão, perca, zoina, puta, rascoa, marafona, bagaxa, dama-da-noite, fuampa, lúrpia, michela, mulher-da-vida, perdida, biraia, rameira, samarrão, pelejo, cróia*

Verbos:

ridicularizar (55 palavras): *embromar, ridicularizar, motejar, trotar, gozar, palhetar, chasquear, malhar, ridiculizar, sotaquear, toirear, chacotear, ironizar, pantear, cachetar, gingar, facetear, gracejar, galhofear, palhetear, desfrutar, debicar, chalacear, mexer, desfruir, zombetear, debochar, achincalhar, pagodear, derriçar, trotear, apepinar, tourear, vasconcear, chalaçar, chocarrear, sorrir, bexigar, mofar, empulhar, cafangar, apodar, troçar, brincar, caçoar, pilheriar, satirizar, rir, bigodear, zingrar, mangar, galhofar, escarnecer, escarnir, zombar*

gritar (53 palavras*): toar, atroar, urrar, exclamar, vozeirar, rebramar, estrilar, troar, tonitruar, estrondar, esbravejar, gritar, trovejar, ulular, uivar, esbravear, bradejar, tempestear, retroar, explodir, bramir, mugir, deblaterar, estrugir, trovoar, vozear, tronar, expluir, esgoelar, ressonar, bravejar, rimbombar, barulhar, conclamar, tempestuar, goelar, barregar, fragorar, ribombar, ressoar, estourar, resoar, vociferar, berrar, retumbar, estrondear, rebombar, bradar, esbravecer, latir, estoirar, berregar, bramar*

enfeitar (50 palavras): *colorear, incrementar, adernar, emoldurar, alfaiar, floretear, formosear, paramentar, assazonar, decorar, tecer, ataviar, aparelhar, enfeitar, enflorear, ornar, sobredoirar, sazonar, aformosear, recamar, amoldurar, aparamentar, brilhantar, sobredourar, embrincar, assazoar, ornamentar, jaezar, emoldar, enflorar, colorar, colorizar, moldurar, aparatar, embelezar, adereçar, alindar, embonecar, engalanar, espenicar, aformosar, honestar, florear, abonecar, adornar, ajaezar, formosentar, galantear, exornar, aformosentar*

O recurso final

Após a integração das relações do PAPEL e das relações extraídas do DA e Wikcionário, obtivemos uma ontologia lexical cujas quantidades de relações, agora entre *synsets*, se apresentam na tabela 7. Apesar de, nesta tabela, não terem sido contabilizadas relações inversas, é sempre possível inferir este tipo de relações, ou seja, quando temos uma relação A hiperónimo-de B, é sempre possível inferir que B hipónimo-de A.

Quando se pretende integrar uma palavra que o *thesaurus* não contém, é necessário criar um novo *synset* apenas com essa palavra. Assim, na tabela 7, optamos por separar as relações de acordo com aquelas que se estabelecem entre dois *synsets* com apenas uma palavra (*1 → 1*), *synset* com uma palavra e *synset* com várias palavras (*1 → n*, *n → 1*) e, ainda, dois *synsets* com mais de uma palavra (*n → n*). Verifica-se que a maior parte das relações liga um *synset* com uma palavra a um *synset* com mais palavras.

Relação	Args	Instâncias			
		1 → n	*1→n ; n→1*	*n→n*	Total
Hiperónimo-de	n,n	2.789	45.637	40.161	88.584
Parte-de	n,n	654	1.910	1.654	4.218
	n,adj	663	2.614	2.017	5.294
	adj,n	23	41	15	79
Membro-de	n,n	776	3.244	2.326	6.346
	n,adj	12	43	13	68
	adj,n	265	366	182	813
Causador-de	n,n	201	580	573	1.354
	n,adj	2	7	12	21
	adj,n	102	221	305	628
	n,v	2	11	41	54
	v,n	388	1.920	4.338	6.646
Produtor-de	n,n	282	627	574	1,473
	n,adj	12	28	24	64
	adj,n	30	224	201	455
Finalidade-de	n,n	620	2.498	2.745	5.863
	n,adj	18	50	33	101
	v,n	1.531	3.617	2.720	7.868
	v,adj	27	183	156	366
Local-origem	n,n	674	520	151	1.345
Maneira-de	adv,n	135	1.073	939	2.147
	adv,adj	47	732	619	1.398
Maneira	adv,n	1	95	137	233
sem	adv,v	1	10	7	18
Propriedade-de	adj,n	1.994	4.181	2.564	8.739
	adj,v	7.761	15.184	2.778	25.723

Tabela 7: Relações entre synsets na versão actual do Onto.PT

Por fim, para dar uma ideia da estrutura e do que é possível encontrar no Onto.PT, mostramos, na figura 3, os resultados da pesquisa pela palavra 'hospital', na versão actual do nosso recurso.

- S: (adj) **hospital**, bem-querente, bem-intencionado, caridoso, benevolente, benévolo
- S: (adj) **hospital**, caridoso, esmoler, benfazejo, bem-fazejo, bem-fazente
- S: (s) **hospital**, esprital, nosocômio
  - o temParte
    - S: (s) enfermaria
      - parteDe
      - hiperonimoDe
        - S: (s) ambulatório
      - hiponimoDe
      - meioPara
        - S: (s) paciente, enfermeiro, doente
  - o referidoPorAlgoComPropriedade
    - S: (adj) hospitalar, hospitalário
    - S: (adj) nosocómico, nosocomial
  - o hiperonimoDe
    - S: (s) gafaria, leprosaria
    - S: (s) leprocómio
    - S: (s) lazareto
    - S: (s) hospital psiquiátrico, manicômio, hospício, rilhafales, casa-de-orates, manicómio
  - o hiponimoDe
    - S: (s) edifício, edificamento, edificação
    - S: (s) instituto, instituição, fundação, instauração, infra-estrutura, estabeleza, implantação, implante, estabelecimento, inauguração
    - S: (s) telhado, mobiliário, teito, tecto, habitação, dinastia, meisom, casa, cosque, mesão

Figura 3: Relações e synsets a partir da entrada 'hospital'.

5. Notas conclusivas

Neste artigo foi apresentada a abordagem adoptada na construção automática de uma ontologia lexical para o português, o Onto.PT, juntamente com um conjunto de exemplos que ilustram tanto as fases de construção como os resultados mais recentemente obtidos.

Um ponto forte deste trabalho é a dimensão dos recursos gerados, onde, ao agregar um conjunto de recursos livres para o português, se inclui a maior rede lexical e também o maior *thesaurus* electrónico para a nossa língua. Além disso, dada a flexibilidade da abordagem seguida, será possível aumentar ainda mais o Onto.PT, tirando partido de outros recursos textuais.

Por outro lado, há ainda um longo caminho a percorrer no que diz respeito à qualidade do recurso e à organização do conhecimento nele

contido. Acreditamos que, brevemente, o Onto.PT seja uma alternativa livre aos recursos lexicais actualmente existentes para o português, e que possa contribuir para um maior desenvolvimento de ferramentas e aplicações dentro do processamento computacional da nossa língua.

Referências

Agirre, Eneko, Oier Lopez De Lacalle, e Aitor Soroa (2009). Knowledge-based WSD on specific domains: performing better than generic supervised WSD. *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI 2009)*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1501-1506.

Amsler, Robert A. (1981). A taxonomy for English nouns and verbs. *Proceedings of the 19th Annual Meeting of The Association for Computational Linguistics*. Morristown, NJ: ACL Press, 133-138.

Chodorow, Martin S., Roy J. Byrd e George E. Heidorn (1985). Extracting semantic hierarchies from a large on-line dictionary. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ: ACL Press, 299-304.

De Melo, Gerard e Gerhard Weikum (2008). On the Utility of Automatically Generated Wordnets. *Proceedings of the 4th Global WordNet Conf. (GWC)*. Szeged, Hungary: University of Szeged, 147-161.

Dias-da-Silva, Bento Carlos, Mirna F. de Oliveira, e Helio R. de Moraes (2002). Groundwork for the Development of the Brazilian Portuguese WordNet. *Proceedigns of Advances in Natural Language Processing, 3rd International Conference (PorTAL),* Faro, Portugal. LNCS 2389, Springer. 189–196.

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

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

Gonçalo Oliveira, Hugo e Paulo Gomes (2011a). Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*. Barcelona, Spain: AAAI Press, 1801–1806.

Gonçalo Oliveira, Hugo e Paulo Gomes (2011b). Ontologising Relational Triples into a Portuguese Thesaurus. *Local Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*. Lisbon, Portugal: APPIA.

Gonçalo Oliveira, Hugo e Paulo Gomes (2011c.) Automatically enriching a Thesaurus with information from Dictionaries. *Progress in Artificial Intelligence. Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, LNCS 7026. Springer, 462–475.

Gonçalo Oliveira, Hugo e Paulo Gomes (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. *Proceedings of the 5th European Starting AI Researcher Symposium (STAIRS 2010)*, Lisbon, Portugal. IOS Press, 199-211.

Gonçalo Oliveira, Hugo, Diana Santos e Paulo Gomes (2010). Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática* 2:1, 77-93.

Hearst, Marti A. (1998). Automated Discovery of WordNet Relations. In Fellbaum, Christiane (ed.) (1998). *WordNet*: *An Electronic Lexical Database* (*Language, Speech, and Communication*). Cambridge, Massachusetts: The MIT Press, 131-151.

Hirst, Graeme (2004). Ontology and the Lexicon. In: S. Staab & R. Studer (eds.) *Handbook on Ontologies*. Springer, 209-230.

Ide, Nancy e Jean Veronis (1995). Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation. In: *Machine Translation and the Lexicon, LNAI*. Springer, 19-34.

Lin, Dekang e Patrick Pantel (2002). Concept Discovery from Text. *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, 577–583.

Marrafa, Palmira (2002). Portuguese WordNet: general architecture and internal semantic relations. *DELTA* 18: 131–146.

Maziero, Erick G., 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. *Proceedings of the VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL),* Vila Velha, Espírito Santo, 390-392.

Medelyan, Olena, David Milne, Catherine Legg, e Ian H. Witten (2009). Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies* 67:9, 716–754.

Nichols Eric, Francis Bond e Dan Flickinger (2005). Robust Ontology Acquisition from Machine-Readable Dictionaries. *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI),* Edinburgh, Scotland, UK. Professional Book Center, 1111–1116.

Pasca, Marius e Sanda M. Harabagiu (2001). The Informative Role of WordNet in Open-Domain Question Answering. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburgh, USA, 138–143.

Pianta, Emanuele, Luisa Bentivogli, e Christian Girardi (2002). Multiwordnet: developing an aligned multilingual database. *1st International Conference on Global WordNet*. Mysore, India, 2002.

Prévot Chu-Ren Huang Laurent, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, e Alessandro Oltramari (2010). Ontology and the Lexicon: a multi-disciplinary perspective (introduction). In: Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, e Laurent Prevot (eds.). *Ontology and the Lexicon: A Natural Language Processing Perspective*. Cambridge: Cambridge University Press, 3–24.

Resnik, Philip (1995). Disambiguating Noun Groupings with Respect to WordNet Senses. *Proceedings of the 3rd Workshop on Very Large Corpora*. Cambridge, MA: The MIT Press, 54-68.

Richardson, Stephen D., William B. Dolan e Lucy Vanderwende (1998). MindNet: Acquiring and Structuring Semantic Information from Text. *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*. ACL, Stroudsburg, PA, USA. 1098–1102.

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, e Rosário Silva (2010). Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In: A. M. Brito, F. Silva, J. Veloso, e A. Fiéis (eds.). *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa: Colibri/APL, 681–700.

Simões Alberto e Rita Farinha (2011). Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa* 16, 159-171.

Teixeira, Jorge, Luís Sarmento e Eugénio Oliveira (2010). Comparing Verb Synonym Resources for Portuguese. *Proceedings of the 9th International Conference* (*PROPOR 2010*), LNAI 6001. Springer, 100-109.

Vossen, Piek (1997). EuroWordNet: a multilingual database for information retrieval. *Proceedings of the DELOS workshop on Cross-Language Information Retrieval*. European Research Consortium For Informatics and Mathematics. Zurich, 85-94.