WordNet.PT* – Uma rede léxico-conceptual do Português on-line

Palmira Marrafa[1,2], *Raquel Amaro*[1], *Rui Pedro Chaves*[1], *Susana Lourosa*[1], *Catarina Martins*[1] *e Sara Mendes*[1]
[1]CLG – Grupo de Computação do Conhecimento Léxico-Gramatical, CLUL
[2]Departamento de Linguística Geral e Românica, Faculdade de Letras
Universidade de Lisboa

1 Introdução

A WordNet.PT (Marrafa (2001) e (2002)) é uma base de dados de conhecimento lexical do Português, que está a ser desenvolvida no quadro da EuroWordNet (Vossen, 1999), uma base de dados multilingue de larga escala que inclui wordnets de várias línguas europeias. A EuroWordNet segue as linhas gerais da WordNet de Princeton, que constitui a primeira base de dados de conhecimento linguístico em que o significado lexical é representado através de uma rede de relações lexicais e conceptuais, sendo o significado de cada unidade lexical derivado da sua posição na rede.

Apresentam-se aqui os fundamentos empíricos, a arquitectura geral e as relações estruturantes da WordNet.PT (WN.PT, de ora em diante), bem como as actuais linhas de orientação. Assim, na secção 2 descreve-se o modelo da WN.PT. Na secção 3 discute-se a problemática da identificação do conjunto de variantes lexicais que correspondem a um mesmo conceito. Na secção 4 explicitam-se as relações de hiponímia/hiperonímia e de meronínima/holonímia. As relações utilizadas para codificar a informação relativa aos participantes típicos de um dado evento são discutidas na secção 5. Na secção 6 são definidas as novas linhas de orientação da WN.PT.

2 Enquadramento

A WN.PT é uma base de dados de conhecimento lexical desenvolvida no âmbito do modelo geral da EuroWordNet, uma base de dados multilingue que integra wordnets de várias línguas europeias.

A organização de cada uma das wordnets assenta essencialmente nos pressupostos da WordNet de Princeton, desenvolvida para o Inglês americano pelo Laboratório de Ciências Cognitivas da Universidade de Princeton (Miller *et al.* (1990), Fellbaum (1998a) e (1998b)). Entende-se, assim, que o léxico mental se encontra estruturado em torno de um conjunto de relações conceptuais de natureza diversa, deduzindo-se o significado das unidades lexicais a partir dessas relações (Miller, 1986).

* Cabe aqui um agradecimento às Instituições que, em diferentes etapas, têm financiado o projecto WordNet.PT: Instituto Camões (1999-2002), Fundação Calouste Gulbenkian (2003-2004) e Fundação para a Ciência e a Tecnologia (2004-2006).

A unidade básica nas wordnets é o conceito. Os conceitos são representados por conjuntos de sinónimos (synsets). Cada *synset* contém todas as lexicalizações de um mesmo conceito e constitui um nó da rede. As expressões *automóvel* e *carro*, por exemplo, estão incluídas no mesmo *synset*, já que ambas são lexicalizações do mesmo conceito.

Ao contrário do que acontece nos dicionários convencionais, nos quais o significado das unidades lexicais é definido por paráfrases, nas wordnets o sentido emerge das relações lexicais e conceptuais que se estabelecem entre elas.

A integração da WN.PT na rede multilingue do modelo da EuroWordNet é feita através do Inter-Lingual-Index (ILI). O ILI é uma lista de conjuntos de sinónimos, correspondentes a conceitos retirados maioritariamente da versão 1.5 da WordNet de Princeton. O ILI permite relacionar as diversas wordnets entre si. Em concreto, cada conjunto de sinónimos[1] de uma língua específica é associado ao conjunto de sinónimos do ILI que representa o mesmo conceito, como o esquema seguinte mostra.

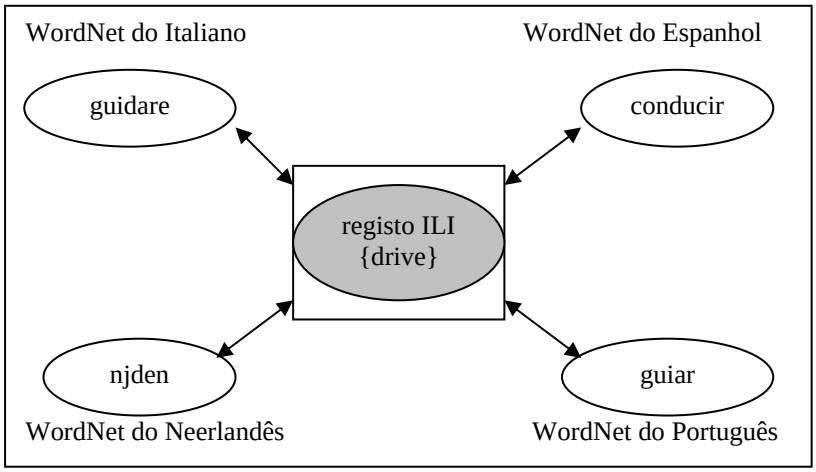

Figura 1. Relações wordnets-ILI.

Nos casos em que não é possível estabelecer uma relação de sinonímia com um synset do ILI, recorre-se a outro tipo de relações, como as de hiperonímia, meronímia e algumas relações funcionais.

3. Identificação e representação dos conceitos

A WordNet.PT está a ser construída manualmente, dado não existirem recursos linguísticos disponíveis que permitam um desenvolvimento automático compatível com o realismo linguístico que se pretende.

[1] Por simplificação, cada conjunto de sinónimos é aqui representado apenas por uma lexicalização.

Dado que a codificação das diferentes categorias envolve a especificação de diferentes relações, optou-se por, numa primeira fase, codificar apenas informação relativa a expressões nominais, tendo-se iniciado a codificação de verbos e adjectivos numa fase subsequente.

A selecção dos conceitos a introduzir é feita por domínios semânticos, o que permite uma cobertura mais exaustiva e incrementa as possibilidades de utilização da rede como base de aplicações diversas.

A determinação das unidades lexicais que integram um *synset* envolve vários níveis de complexidade, que vão da definição de sinonímia à diferenciação *vs* colapsamento de conceitos em caso de polissemia.

A sinonímia é uma relação essencial para a construção de wordnets já que expressões sinónimas integram o mesmo synset, representando, por conseguinte, o mesmo conceito.

Considerando que a verdadeira sinonímia é rara, se é que existe, as wordnets usam uma definição mais fraca, a definição de sinonímia relativa a um contexto (Miller *et. al*, 1990):

Se A é sinónimo de B em C, então B é sinónimo de A em C:

(1) a. Um automóvel é um carro. Verdadeiro
b. Um carro é um automóvel. Verdadeiro
c. Um táxi é um automóvel. Verdadeiro
d. Um automóvel é um táxi. Falso

Face à definição assumida, os dados em (1) evidenciam que *carro* e *automóvel* se encontram no mesmo *synset*, enquanto *táxi* pertence a um *synset* diferente.

Uma das principais dificuldades que se colocam aquando da identificação dos conceitos, e, consequentemente, dos *synsets*, diz respeito aos casos de polissemia. Há que evitar quer a sobrediferenciação quer a subdiferenciação de sentidos, o que não é uma questão trivial, dado que, por um lado, a polissemia não é um fenómeno uniforme, e, por outro, a pertinência dos níveis de diferenciação, depende, em muitos casos, das aplicações em que as wordnets vierem a ser utilizadas. No caso da WN.PT, procura-se uma abordagem equilibrada, i.e., linguisticamente motivada e susceptível de servir o maior número possível de aplicações. Tem-se crucialmente em conta a avaliação do grau de discretude dos diferentes sentidos, ou facetas de sentido (Cruse, 2000), associados a uma dada unidade lexical. Os critérios adoptados serão aferidos, e poderão sofrer alguma alteração, em função dos resultados da utilização da WN.PT no desenvolvimento de projectos colaterais, como os que envolvem sistemas de pergunta-resposta (Marrafa *et al.* (2004)) e de tradução automática (Marrafa *et al.* (2005)).

4 Relações hierárquicas

As relações estruturantes da WN.PT são, fundamentalmente, as relações de hiponímia/hiperonímia e de meronímia/holonímia.

A hiponímia/hiperonímia é uma relação hierárquica fundamental na estruturação das wordnets. Trata-se de uma relação inversa, assimétrica e transitiva, como abaixo se define, informalmente:

i) A é um hipónimo de B
se
A é um tipo de B
B não é um tipo de A

ii) B é um hiperónimo de A
se A é um hipónimo de B

Considerem-se os seguintes exemplos:

(2) a. Um táxi é um tipo de automóvel. Verdadeiro
b. Um automóvel é um tipo de táxi. Falso

(3) a. # O João comprou um táxi, mas não comprou um automóvel.
b. O João comprou um automóvel, mas não comprou um táxi (comprou um comercial).

Os exemplos evidenciam uma maior especificidade de *táxi* relativamente a *automóvel*. Assim, *táxi* tem todas as propriedades de *automóvel*, mas *automóvel* não tem todas as propriedades de *táxi*. Tendo em conta a definição acima apresentada, *táxi* é hipónimo de *automóvel* e *automóvel* é hiperónimo de *táxi*.

Cada *synset* tem associada uma glosa que inclui o hiperónimo imediato e as propriedades específicas do conceito. Este formato permite detectar especificações inadequadas da relação de hiponímia/hiperonímia.

Ainda no que diz respeito às relações de hiponímia/hiperonímia, dá-se especial relevo às estratégias de determinação de hiponímia genuína, prevenindo o estabelecimento de falsa co-hiponímia.

Quanto às relações de meronímia/holonímia, estas oferecem grande resistência a um tratamento formal, enquanto relação uniforme. Com efeito, como tem sido sublinhado por vários autores, trata-se antes de uma complexa família de relações (cf. Cruse, 1986). Canonicamente define-se a relação de meronímia/holonímia e a sua inversa como se segue:

A é merónimo de B se e só se A é necessariamente parte de B
e B inclui necessariamente A;
B é holónimo de A se e só se A é merónimo de B.

É esta a relação que se estabelece entre *pétala* e *corola*. *Pétala* é merónimo canónico de *corola* e *corola* é holónimo canónico de *pétala*, uma vez que *pétala* é necessariamente parte de *corola* e *corola* inclui necessariamente *pétala*.

No entanto, a relação de meronímia/holonímia pode ser facultativa num dos sentidos, ou em ambos. Esse carácter não canónico da relação é explicitamente assinalado na WN.PT através de um traço de reversibilidade, disponível para todas as relações bidirecionais.

A natureza deste tipo de relações não é captada a um nível de granularidade satisfatório se tratada estritamente como relação de inclusão, como geralmente acontece, já que as partes apresentam diferentes graus e formas de integração no todo. Na WN.PT, à semelhança do que acontece na EuroWordNet, distinguem-se cinco subtipos de relações parte/todo:

mero_parte	pétala/corola
mero_membro	futebolista/equipa
mero_porção	talhada/melão
mero_matéria	algodão/sarja
mero_localização	Lisboa/Portugal

Todos estes subtipos poderiam ser acomodados na relação subespecificada, meronímia/holonímia, como acontece na WordNet de Princeton. Contudo, só se opta pela subespecificação nos casos em que não há motivação suficiente para optar por um dos subtipos. Esta subtipificação potencia as possibilidades de exploração da rede, em particular em aplicações que envolvam sistemas de inferência ou de pesquisa de informação.

5 Relações de função

As relações de função são utilizadas para codificar informação relativa aos participantes típicos de um dado evento. Embora menos estruturantes do que as relações referidas no ponto anterior, têm forte motivação cognitiva e revelam-se de grande interesse em aplicações de apoio ao ensino das línguas, por exemplo.

Estas relações permitem estabelecer relações transcategoriais, como mostra o esquema seguinte:

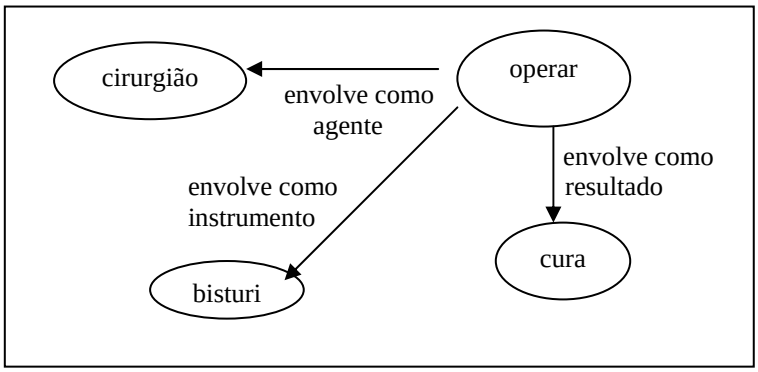

Fig.2. Relações de função na WN.PT

6 Novas directrizes na WordNet.PT

A integração de expressões adjectivais e de certas expressões verbais na WN.PT, veio a revelar necessária a introdução de novas relações, em particular de relações transcategoriais.

Os adjectivos descritivos, por exemplo, que ligam um valor de um atributo a um nome, são relacionados na WN.PT com o atributo que modificam através da relação *caracteriza relativamente a /é caracterizado relativamente a*. Esta relação permite codificar de forma integrada a classe de adjectivos em questão, já que todos os adjectivos que modificam o mesmo atributo são relacionados com o nome que o lexicaliza (cf. Mendes (2006)), como ilustrado na Fig. 3.

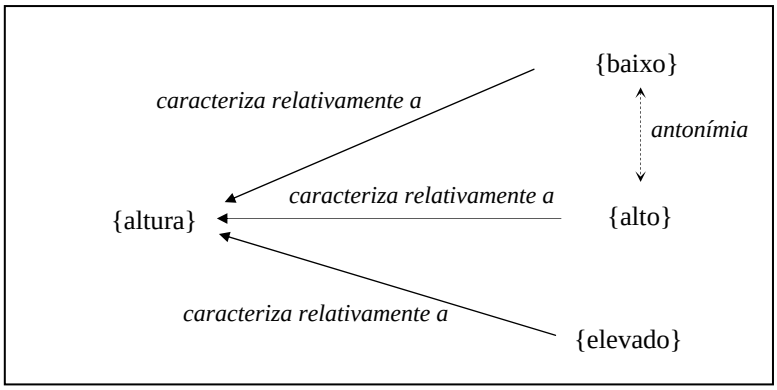

Fig.3. Relações em torno da lexicalização do atributo *altura* na WN.PT

Em relação à codificação de verbos que denotam transições, mas que são deficitários no que respeita às propriedades do estado final do evento, como *tornar*, por exemplo, e cujos tropónimos geralmente incorporam o referido estado final (cf.

entristecer ≡ *tornar triste*), é fortemente motivada a inclusão de uma relação relacionada com a estrutura dos eventos: *tem como subevento télico/é_subvento télico de*.

Deste modo, preservando a coerência do modelo, a WN.PT orienta-se nesta fase para especificações mais ricas, que incluem, entre outra, informação sobre a estrutura argumental e a estrutura dos eventos, e envolvem vários tipos de relações transcategoriais.

Em consequência, além de se conferir à rede um carácter mais integrado, incrementam-se as suas potencialidades no que respeita à sua exploração em aplicações no âmbito da Linguística Computacional e da Engenharia da Linguagem.

Referências

Cruse, D. A. (1986) *Lexical Semantics*.Cambridge: Cambridge University Press.

Cruse, D. A. (2000) Aspects of the Micro-structure of Word Meanings. In Y. Ravin and C. Leacock (eds.) *Polysemy: Theoretical and Computacional Approaches.* Oxford: Oxford University Press.

Fellbaum, C. (1998a) A Semantic Network of English Verbs. In C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Fellbaum, C. (1998b) A Semantic Network of English: The Mother of All WordNets. In P. Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers, pp. 209-220.

Marrafa, P. (2001) *WordNet do Português: uma base de dados de conhecimento linguístico*. Lisboa: Instituto Camões.

Marrafa, P. (2002) Portuguese WordNet: General Architecture and Internal Semantic Relations. *D.E.L.T.A*, 18, pp. 131-146.

Marrafa, P., C. Ribeiro e R. Santos (2005) Automatização da Geração de Dicionários Tratáveis por Máquina: Reutilização de Recursos Linguísticos, *Revista Iberoamericana de Sistemas, Cibernética e Informática*.

Marrafa, P., C. Ribeiro, R. Santos e J. Correia (2004) Gathering Information from a Relational Lexical-Conceptual Database: A Natural Language Question-Answering System. In *Proc. of The 8th World Multi-Conference on Systemics, Cybernetics and Informatics*. Orlando: USA.

Marrafa, P. e S. Mendes (2006) Increasing WordNets Expressive Power and Linguistic Adequacy, artigo submetido ao *PROPOR 2006*.

Mendes, S. (2006) Adjectives in WordNet.PT. In *Proc. of the GWC 2006 – Global WordNet Association Conference*. Jeju Island: Korea.

Miller, G. (1986) Dictionaries in Mind, *Language and Cognitive Processes*, 1, pp. 171-185.

Miller, G. et. al (1990) Introduction to WordNet: An On-line lexical Database. *International Journal of Lexicography*, 3(4), pp. 235-244.

Vossen, P. (1999) *EuroWordNet: General Document*. Universidade de Amesterdão.