Proposta de Classificação Semântica de Unidades Lexicais Multipalavra Nominais

Silvana Abalada[1], Vera Cabarrão[1], Aida Cardoso[1]
[1]Faculdade de Letras da Universidade de Lisboa
silvanaabalada@gmail.com, veracabarrao@gmail.com, aidacard@gmail.com

Introdução

As Unidades Lexicais Multipalavra (ULM) são estruturas linguísticas cristalizadas que têm um peso considerável no conteúdo informativo de qualquer tipo de texto (Ranchhod & Carvalho, 2003). Estas têm vindo, assim, a merecer destaque em aplicações automáticas na área do Processamento de Língua Natural (PLN). Face a extensas bases de dados e a perguntas selectivas de um utilizador, identificar e classificar estas estruturas é, desta forma, essencial para a recuperação de informação (Bick, 2006).

Embora existam já recursos disponíveis, as questões de terminologia e anotação de colecções douradas são dois dos principais problemas nesta área.

Salientem-se de entre os diversos estudos, a nível internacional, por um lado, o léxico semântico de Lancaster e, por outro, o WordNet e o EuroWordNet, constituindo os citados léxicos dois tipos de abordagens possíveis. Já a nível nacional, evidenciem-se o léxico TemaNet e o etiquetador EELO.

Objectivo

Propor uma classificação semântica de ULM nominais para Português Europeu, cujo objectivo é a criação de uma tipologia classificativa adequada a textos pertencentes a domínios gerais e adaptável a diferentes *corpora.*

Metodologia

Corpus CETEMPúblico: extractos de artigos do jornal *Público* recolhidos entre 1991 e 1998 (Rocha & Santos, 2000).

1.º Processamento automático do *corpus,* realizado pelo sistema Unitex, de modo a extrair uma lista de ULM;

2.º Tratamento manual da lista para excluir Entidades Mencionadas (EM) e as ULM não nominais;

3.º Lematização manual das formas flexionadas de ULM nominais e uniformização de ocorrências de palavras com dupla grafia;

4.º Etiquetagem da colecção dourada de ULM nominais (constituída por 5068 itens) com a classificação semântica proposta, sendo que uma amostra aleatória de 507 ULM (10% do *corpus*) foi etiquetada separadamente por três anotadoras com experiência linguística, de modo a validar a aplicação da classificação.

Classificação Semântica Proposta

A classificação semântica proposta, alicerçada no trabalho de Piao *et alii* (2005), foi estruturada como um *thesaurus* e hierarquizada em classes e subclasses (cf. Tabela 1), de forma a garantir uma forte relação semântica (Jurafsky & Martin, 2008).

Classes	Subclasses	
Indivíduo e Corpo Humano (IN)	Identificação (IN1)	
	Vestuário, Acessórios e Cosmética (IN2)	
	Medicina (IN3)	Anatomia e Fisiologia (IN3.1)
		Saúde e Tratamentos Médicos (IN3.2)
Sociedade (SO)	Vida Profissional (SO1)	Mundo do Trabalho (SO1.1)
		Profissões, Cargos e Actividades (SO1.2)
		Organizações, Instituições e Empresas (SO1.3)
	Relações Interpessoais (SO2)	
	Vida em Comunidade (SO3)	
	Estatutos, Grupos e Filiação (SO4)	
	Eventos (SO5)	
Governo e Domínio Público (GO)	Governo e Geopolítica (GO1)	
	Jurisdição e Justiça (GO2)	
	Guerra, Defesa, Exército e Armas (GO3)	
Economia (EC)	Economia, Finanças e Dinheiro (EC1)	
	Comércio, Indústria e Serviços (EC2)	
	Moeda (EC3)	
Arquitectura e Design (AR)	Infra-estruturas (AR1)	
	Habitação e Edifícios (AR2)	
	Partes da Casa e Mobiliário (AR3)	
	Ferramentas e Utensílios (AR4)	
	Equipamentos e Electrodomésticos (AR5)	
Localização e Movimento (LO)	Vias e Meios de Transporte (LO1)	
	Navegação e Circulação (LO2)	
	Local e Espaço (LO3)	
Gastronomia (GA)	Produtos Primários (GA1)	
	Produtos Secundários (GA2)	
Mundo Animal e Vegetal (MU)	Animais (MU1)	
	Plantas (MU2)	
Ambiente (AM)	Meio Ambiente e Recursos Energéticos (AM1)	
Educação (ED)	Geral (ED1)	
Linguística (LI)	Geral (LI1)	
Comunicação (CM)	Comunicação Humana (CM1)	
	Comunicação Social (CM2)	
Ciência e Conhecimento (CI)	Ciências Exactas e Tecnologia (CI1)	
	Ciências Sociais e Humanas (CI2)	
Cultura, Entretenimento e Desporto (CU)	Cultura, Artes e Espectáculos (CU1)	
	Desporto e Jogos (CU2)	
Astronomia (AS)	Geral (AS1)	
Esoterismo e Religião (ES)	Esoterismo (ES1)	
	Religião (ES2)	
Tempo (TE)	Geral (TE1)	
	Período (TE2)	
	Idade (TE3)	
Matérias e Substâncias (MA)	Geral (MA1)	
Medidas (MD)	Geral (MD1)	
Cores (CR)	Geral (CR1)	
Disposições Emocionais, Atitudes e Comportamentos (DI)	Geral (DI1)	
Avaliação e Validação (AV)	Geral (AV1)	
Conceitos (CO)	Geral (CO1)	
Metáforas (ME)	Geral (ME1)	

Tabela 1: Proposta de Classificação Semântica de ULM Nominais para Português Europeu

Resultados

A Classificação Semântica permitiu etiquetar 97% do *corpus* (das 5068 ULM, 147 não foram classificadas).

A aplicação da Classificação Semântica foi validada, já que se verificou um acordo entre as anotadoras em 96% dos casos (das 507 ULM, 486 foram etiquetadas com acordo).

A distribuição absoluta de ULM por classe reflecte a natureza do *corpus,* ou seja, os domínios típicos de um jornal generalista, representados, nomeadamente, pelas classes Sociedade e Governo e Domínio Público, têm um maior predomínio (cf. Gráfico 1).

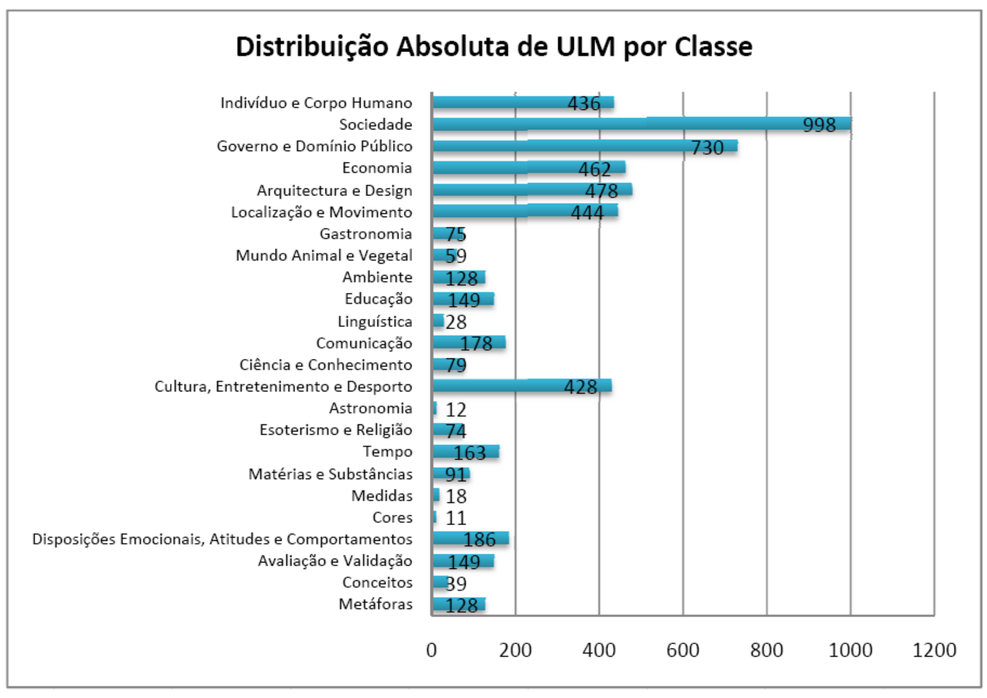

Gráfico 1: Distribuição Absoluta das ULM por Classe

A análise do *corpus* conduziu a uma etiquetagem não única das ULM, como ilustra o gráfico da distribuição de ULM por número de etiquetas (cf. Gráfico 2). Ainda assim, verifica-se que, na maioria dos casos, a classificação única revelou-se a mais produtiva, o que comprova a eficácia da classificação semântica .

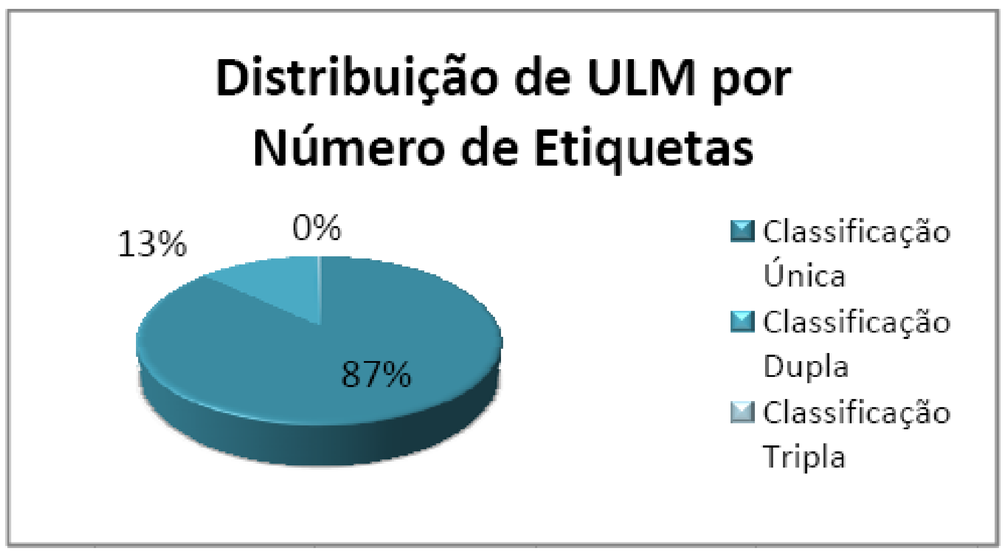

Gráfico 2: Distribuição de ULM por Número de Etiquetas

Conclusões

Este estudo, ainda que preliminar e exploratório, pretendeu aproximar-se da dimensão da colecção dourada proposta pelo HAREM (cerca de 5270 EM). Apesar de o nosso trabalho tratar estruturas linguísticas distintas, esta aproximação revela-se pertinente na medida em que se colocam o mesmo tipo de problemas e a possibilidade de usar a colecção dourada aqui proposta em futuras avaliações na área do PLN.

Aqui apresentou-se uma proposta de classificação semântica de ULM nominais para PE que se pretende que seja um ponto de partida para futuros estudos nesta área. A necessidade latente de desenvolver léxicos mais completos, ou seja, tratados não só a nível morfossintáctico, mas igualmente semântico, justifica tal expectativa.

Finalmente, considera-se que, com este estudo, não só se estabeleceu uma linha de continuidade com trabalhos realizados para outras línguas, que fazem igualmente uso da proposta do léxico semântico de Lancaster, como se trouxe para a discussão uma proposta diferente das existentes a nível nacional e, por conseguinte, algo inovador em PE, no âmbito do PLN.

Agradecimentos

Agradecemos à Professora Doutora Elisabete Ranchhod, uma vez que o presente trabalho foi inicialmente desenvolvido no âmbito do seminário de Linguística Computacional: Processamento das Línguas Naturais, do Mestrado em Linguística, da FLUL; ao Professor Doutor Nuno Mamede, por nos ter facultado o acesso à totalidade do *corpus*; à Helena Moniz, pelos tão preciosos comentários e críticas e ao José Portêlo pelo apoio técnico.

Bibliografia

Bick, E. (2006): *Noun Sense Tagging: Semantic Prototype Annotation of Portuguese Treebank*, *in* Hajic, J. & J. Nivre (eds.): *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, Praga.

Jurafsky, D. & J. H. Martin (2008): *Speech and Language Processing: An Introduction to Natural Languages Processing, Speech Recognition, and Computational Linguistics*, New Jersey, Prentice-Hall.

Mota, C.; Santos, D. & Ranchhod, E. (2007): "Avaliação de reconhecimento de entidades mencionadas: princípio de HAREM", *in* Santos, D. (ed.): *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, IST-Press, Lisboa.

Mota, C. (2009): *How to keep up with Language Dynamics: A case-study on Named Entity Recognition,* Tese de Doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa.

Piao, Scott S. L. *et alii* (2005): "A Large Semantic Lexicon for Corpus Annotation", *in Proceedings from The Corpus Linguistics Conference Series, Corpus Linguistics 2005*, Birmingham.

Ranchhod, E. & P. Carvalho (2003): "Unidades Lexicais Complexas. Problemas de Análise e Etiquetagem", *in Actas do VIII Simpósio International de Comuicación Social*, Santiago de Cuba.

Rocha, P. A. & D. Santos (2000): "CETEMPúblico: Um *corpus* de grandes dimensões de linguagem jornalística portuguesa", *in* Nunes, Maria das Graças (ed.): *Actas do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Atibaia.