Proposta de Classificação Semântica de Unidades Lexicais Multipalavra Nominais[1]

Silvana Abalada, Vera Cabarrão & Aida Cardoso
Faculdade de Letras da Universidade de Lisboa

Abstract

We present a semantic taxonomy to classify nominal multiword lexical units (MLU) for European Portuguese. Despite being built by single words, MLU don't have a compositional meaning and have morphosyntactic restrictions. These units are so important in any text that their identification and classification is essential for information extraction and retrieval in Natural Language Processing. We adapted and applied a semantic taxonomy, based on the Lancaster semantic lexicon (Piao *et alii*, 2005), to a list of MLU extracted from CETEMPúblico. The results of the annotation task validated our taxonomy, because we were able to classify 97,1% of the corpus.

Keywords: Semantic Taxonomy, Multiword Lexical Units, Natural Language Processing.
Palavras-chave: Classificação Semântica, Unidades Lexicais Multipalavra, Processamento de Língua Natural.

1. Introdução

O presente trabalho consiste numa proposta de classificação semântica de unidades lexicais multipalavra (ULM) nominais para Português Europeu. A proposta compreende dois aspectos: a adaptação de uma taxonomia semântica e a aplicação da mesma a um *corpus*.

Embora não tenham sido tratadas enquanto tal na tradição gramatical e na gramática generativa, as ULM nominais aproximam-se, de forma mais ou menos explícita, a alguns itens linguísticos descritos naquelas gramáticas. Quanto à tradição gramatical,

[1] Este artigo é resultado da apresentação de um *poster* no *XXV Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, 22 a 24 de Outubro de 2009.

***Textos Seleccionados, XXV Encontro Nacional da Associação Portuguesa de Linguística,* Porto, APL, 2010, pp. 81-94**

as ULM aproximam-se às palavras compostas formadas por justaposição, pois ambas representam "sempre uma ideia única e autónoma, muitas vezes dissociada das noções expressas pelos seus componentes." (Cunha & Cintra, 2000: 106). Já quanto à gramática generativa, as ULM podem ser associadas a dois tipos de itens linguísticos: os "compostos morfossintácticos" e as "expressões sintácticas lexicalizadas" (Villalva, 2003).

O termo ULM surge, pois, no âmbito do Processamento de Língua Natural (PLN) para nomear expressões linguísticas cristalizadas, como, por exemplo, "bilhete de identidade", "meia de seda", "colo do útero", "artrite reumatóide" e "curso de mestrado". Por expressões linguísticas cristalizadas, especificamente nominais, entende-se estruturas que se distinguem das unidades lexicais simples por apresentarem, do ponto de vista semântico, uma interpretação não composicional (por exemplo, "bode expiatório") e por terem, do ponto de vista morfossintáctico, restrições de flexão (por exemplo, "boas maneiras"), sendo que a reforçar estes dois aspectos está o facto de, em contexto interno de ULM, não ser possível inserir material lexical. Refira-se, ainda, que uma expressão linguística para ser identificada como ULM não tem necessariamente de responder de forma positiva a todos os critérios.

Tendo em conta que as ULM têm um peso considerável no conteúdo informativo de qualquer tipo de texto (Ranchhod & Carvalho, 2003), o estudo das mesmas tem vindo a merecer destaque em aplicações automáticas na área do PLN. Face a extensas bases de dados e a perguntas selectivas de um utilizador, identificar e classificar estas estruturas é essencial para a extracção e recuperação automáticas de informação de *corpora* (Bick, 2006). Embora já existam recursos disponíveis, as questões de terminologia e anotação de *corpora* são dois dos principais problemas no PLN. Salientem-se de entre os diversos estudos, a nível internacional, por um lado, o léxico semântico de Lancaster[2] e, por outro, a WordNet[3], a EuroWordNet[4] e a MultiWordNet[5], pois estes constituem dois tipos de abordagem possíveis. Já a nível nacional, evidenciem-se a rede léxico-conceptual TemaNet[6], a base de dados lexical MultiWordnet of Portuguese[7] e o etiquetador EELO[8].

Deste modo, a proposta de classificação semântica aqui em causa surge na sequência do crescente interesse multidisciplinar em torno das ULM. Refira-se que a proposta procura contribuir para, primeiro, o incremento do estudo das ULM do ponto de vista semântico; segundo, a criação de uma taxonomia adequada a textos pertencentes a domínios gerais e adaptável a diferentes *corpora*; e, terceiro, uma extracção e recuperação automáticas de informação mais adequadas e eficazes que permitam aos utilizadores um

[2] http://ucrel.lancs.ac.uk/usas/.
[3] http://wordnet.princeton.edu/.
[4] http://www.illc.uva.nl/EuroWordNet/.
[5] http://multiwordnet.fbk.eu/english/home.php.
[6] http://www.instituto-camoes.pt/temanet/inicio.html.
[7] http://lxcenter.di.fc.ul.pt/services/pt/LXServicesWordnetPT.html.
[8] http://label.ist.utl.pt/pt/eelo_intr_pt.php.

acesso mais fácil a extensas bases de dados. Além disso, e porque esta proposta aplica uma metodologia já testada em *corpora* de outras línguas, é possível uma comparação dos resultados da sua aplicação.

O presente artigo está organizado em cinco partes. Após esta introdução do tema e da motivação deste trabalho, na segunda parte, apresenta-se o enquadramento teórico. Na terceira, descreve-se o *corpus* adoptado. Na quarta, explicita-se o processo de adaptação da classificação semântica e subsequente aplicação. Por último, tecem-se algumas considerações finais, enunciando hipóteses de trabalho futuro.

2. Enquadramento Teórico

2.1. Definição e Identificação de ULM

Nos últimos anos, as ULM têm vindo a ser alvo de um estudo mais sistemático, contrariando assim o lugar marginal a que estas eram votadas na tradição gramatical. Não obstante o trabalho já realizado, a correcta identificação e posterior classificação das ULM apresenta ainda dificuldades, muito devido ao facto de a sua definição não ser totalmente consensual.

Quanto à descrição e identificação das ULM, importa, antes de mais, ter em consideração que estas, por serem expressões cristalizadas, podem ser relacionadas com outras estruturas linguísticas definidas, de acordo com o *idiom principle* (Sinclair, 1991), como estruturas previamente concatenadas de que os falantes dispõem a nível do processamento.

Ainda assim, um dos problemas reside na descrição e identificação das ULM por oposição às combinações livres de palavras simples. Pese embora o facto de tal distinção constituir uma dificuldade teórica, tentativas várias têm procurado afastar-se da tradição gramatical e lexicográfica através da definição de critérios puramente linguísticos que permitam definir de modo claro estes itens linguísticos e, consequentemente, identificá--los correctamente.

Neste contexto, a noção de fixidez desempenha um papel central (Ranchhod, 2003). Gross (1988), Baptista (1995) e Ranchhod (2001) propõem, pois, a utilização de "um conjunto de critérios linguísticos, que vão desde a verificação do comportamento morfológico (restrições sobre a flexão) dos seus constituintes até à verificação da sua, total ou parcial, perda de composicionalidade lexical, sintáctica e semântica" (Ranchhod, 2003: 4).

O estudo da estrutura interna das ULM revela-se também importante no que respeita à identificação da categoria morfossintáctica de cada palavra simples que constitui estas unidades e às relações que as palavras simples estabelecem entre si. Tal estudo é indispensável na determinação do núcleo de uma ULM e, consequentemente, da sua categoria, condição *sine qua non* para uma correcta análise linguística destes itens e, assim sendo, uma correcta identificação e subsequente classificação morfológica, sintáctica e semântica.

2.2. Classificação de ULM

Tal como a identificação, a classificação de ULM, seja esta morfológica, sintáctica ou semântica, continua a apresentar algumas dificuldades. Todavia, e quanto à análise semântica, âmbito de estudo deste trabalho, devem destacar-se alguns estudos que incidem sobre o tratamento de ULM, mas também de palavras simples.

A nível internacional, refiram-se o léxico semântico de Lancaster, por um lado, e a WordNet, a EuroWordNet e a MultiWordNet, por outro. O léxico semântico de Lancaster (Piao *et alii*, 2005), no qual se baseia este trabalho, consiste num conjunto de classes semânticas organizadas como um *thesaurus*. Já a WordNet (Fellbaum, 1998, *apud* Piao *et alii*, 2005), a EuroWordNet (Vossen, 1998, *apud* Piao *et alii*, 2005) e a MultiWordNet (Pianta, Bentivogli & Girardi, 2002) apresentam as palavras agrupadas por relações entre sentidos. Assim sendo, enquanto no léxico semântico de Lancaster é proposta uma concepção do mundo tão geral quanto possível (uma análise de conteúdo), na WordNet, na EuroWordNet e na MultiWordNet estabelece-se uma rede semântica construída para domínios específicos (uma análise conceptual).

Por seu turno, a nível nacional, refiram-se o etiquetador EELO (desenvolvido pelo LabEL: Laboratório de Engenharia da Linguagem), a MultiWordnet of Portuguese (desenvolvido pelo NLX: Natural Language and Speech Group) e a rede léxico-conceptual TemaNet (Marrafa, 2001). O primeiro é, segundo a equipa do LabEL, um etiquetador linguístico de textos que inclui léxicos e algumas gramáticas de resolução de ambiguidades. Assinale-se que a etiquetagem linguística é morfossintáctica, sendo, por vezes, acompanhada de notações semânticas respeitantes a alguns domínios (como, por exemplo, "Botânica", "Cargos", "Culinária", "Segurança" e "Vestuário"). O segundo, sendo um projecto baseado no modelo da WordNet, tem como objectivo estabelecer relações semânticas de hiperonímia e hiponímia entre palavras, incluindo, como referem os autores, sub-ontologias subordinadas aos conceitos "Pessoa", "Organização", "Evento", "Localização" e "Artes". O terceiro, sendo igualmente um projecto baseado no modelo da WordNet, visa, de acordo com a autora, construir *wordnets* organizadas em domínios semânticos (como, por exemplo, "Alimentação", "Comunicação", "Saúde", "Transportes" e "Vestuário"), ou seja, redes léxico-conceptuais que incluem expressões lexicais ligadas entre si por relações de sinonímia, hiperonímia, hiponímia, holonímia e meronímia.

3. Descrição do *Corpus*

A escolha do *corpus* no qual se baseia este trabalho procurou ter em conta que este teria de servir dois fins: a adaptação da classificação semântica e a sua posterior aplicação.

Neste sentido, o *corpus* seleccionado foi o CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público)[9]. Este *corpus* é constituído por um conjunto de ex-

[9] http://www.linguateca.pt/cetempublico/.

tractos de artigos do jornal *Público* recolhidos entre 1991 e 1998 (Rocha & Santos, 2000) no âmbito de um projecto de investigação de Processamento Computacional do Português, financiado pelo governo, e representa uma das maiores bases de dados do Português Europeu escrito. Refira-se, ainda, que o CETEMPúblico é composto por vinte partes, cada uma com 80.000 extractos, à excepção da vigésima que contém 37.500 extractos. A respeito do *corpus*, saliente-se que o seu conteúdo não representa um domínio específico, cobrindo sim os domínios típicos de um diário generalista em Portugal. Por outras palavras, são privilegiadas áreas como a Política, a Sociedade, a Economia e o Desporto, em detrimento de outras como, por exemplo, a Filosofia, a Física ou a Química. Finalmente, refira-se que, sendo o CETEMPúblico de acesso público, é possível realizar análises mais aprofundadas e comparativas de dados provenientes de vários trabalhos.

Note-se que, tendo em conta que a classificação semântica aqui proposta incide exclusivamente sobre a análise de ULM nominais, foi necessário proceder a um tratamento prévio do *corpus* para obter uma lista de ULM, o que implicou, por sua vez, a adopção de uma metodologia faseada.

Na primeira fase, processou-se automaticamente a totalidade do *corpus* recorrendo ao sistema Unitex[10] (desenvolvido pelo LADL: Laboratoire d'Automatique Documentaire et Linguistique, mas fazendo uso do léxico do LabEL: LABEL-LEX), de modo a extrair todas as ocorrências de ULM, resultando daí uma lista de 161.350 ULM. Na segunda, procedeu-se ao tratamento manual da lista para excluir ULM não nominais e entidades mencionadas. Estas últimas, embora sejam ULM, foram excluídas por requererem um tratamento específico, já efectuado no âmbito de outros trabalhos, nomeadamente dos trabalhos de avaliação dos sistemas de reconhecimento de entidades mencionadas: HAREM (Mota, Santos & Ranchhod, 2007). Na terceira, e última fase, eliminaram-se da lista extraída automaticamente ULM repetidas, de forma a contabilizar apenas uma ocorrência lematizada de cada ULM. Advirta-se para o facto de, entre as ULM repetidas, se terem registado casos de dupla grafia, optando-se, neste caso, por apenas contar uma das ocorrências, sendo esta seleccionada com base nas normas definidas por um prontuário do Português Europeu (Bergström & Reis, 2007). Do tratamento do *corpus* obteve-se, pois, uma lista com 5068 ULM nominais.

Considerando os dados de estudos anteriores (Mota, Carvalho & Ranchhod, 2004) e a discrepância entre o valor de ULM inicialmente extraído pelo Unitex (161.350) e o valor final de ULM nominais (5068), poder-se-ia pensar que as ULM teriam uma frequência baixa. Porém, estas diferenças dever-se-ão em muito à adopção de diferentes métodos de extracção de ULM e à eliminação de todas as repetições e não à sua baixa frequência. Neste sentido, a lista de ULM nominais obtida, pelo seu número de ocorrências, aproxima--se da colecção dourada proposta pelo HAREM (Mota, Santos & Ranchhod, 2007), pelo que se poderá colocar a hipótese de a mesma vir a ser considerada uma colecção dourada em futuras avaliações de trabalhos sobre ULM.

[10] http://www-igm.univ-mlv.fr/~unitex/.

Por último, note-se que na lista final de ULM não foram contempladas informações morfossintácticas e semânticas provenientes da extracção automática do Unitex, dado que não cabia aqui realizar uma análise morfossintáctica e que se pretendia aplicar a proposta aqui apresentada sem conflitos entre diferentes taxonomias.

4. Proposta de Classificação Semântica

4.1. Adaptação da Classificação

Definido o *corpus*, a adaptação da proposta de classificação semântica foi realizada em dois momentos distintos. O primeiro corresponde ao estabelecimento de critérios de adaptação da classificação de Piao *et alii* (2005), tendo este sido efectuado com base no conhecimento das características do *corpus*. Já o segundo diz respeito ao estabelecimento de critérios de anotação, tendo este decorrido mediante uma primeira observação dos dados.

4.1.1. Critérios de Adaptação

Tendo sido baseada no trabalho de Piao *et alii* (2005), a classificação semântica proposta encontra-se estruturada como um *thesaurus*, segundo uma hierarquia de classes e subclasses, de forma a garantir uma forte relação semântica baseada nos conceitos de sinonímia, antonímia, hiperonímia, hiponímia, holonímia e meronímia, tal como também defendem, por exemplo, Jurafsky & Martin (2008).

A opção de basear este trabalho no léxico semântico de Lancaster (Piao *et alii*, 2005) justifica-se pelo facto de um dos objectivos ser a proposta de uma classificação estruturada como um *thesaurus*. Por sua vez, a preferência pela escolha de uma organização em *thesaurus* deve-se ao facto de se pretender uma análise geral do conteúdo informativo do material linguístico. Saliente-se que este tipo de análise apresenta duas grandes vantagens.

A primeira vantagem prende-se com o facto de o material linguístico se encontrar organizado por campos semânticos e não por relações entre palavra e significado, tal como os autores explicitam.

> *The Lancaster semantic lexicon classifies lexemes under a set of broadly defined semantic field categories such as "food and farming", "Life and living things", etc., which are organised, in turn, in a thesaurus-like structure (cf. WordNet and EuroWordNet, in which lexemes are clustered and linked via the relationship between word/MWE senses or definitions of meaning). While word sense indisputably provides the substantial information for linking and organising words, the semantic field (or lexical field) identifies "named area[s] of meaning in which lexemes interrelate and define each other in specific ways" (Crystal 1995) and, as such, has long been used as a framework for structuring lexemes (…).* (Piao *et alii*, 2005: 2)

A referida organização por campos semânticos é conseguida por uma estrutura de classes (respeitantes às áreas gerais), sendo que estas podem ainda ser constituídas por subclasses (estas respeitantes às especificações semânticas das áreas gerais)[11]. Note-se que, dentro de cada classe, o material linguístico mantém entre si relações semânticas (sinonímia, antonímia, hiperonímia, hiponímia, holonímia e meronímia), o que confere homogeneidade às classes.

A segunda vantagem, por seu turno, relaciona-se com o facto de esta abordagem, por ser geral, permitir a aplicação da classificação a *corpora* mais abrangentes, mas também mais específicos, o que não seria viável com uma taxonomia para domínios específicos. A título de exemplo, mencione-se que, com esta classificação, seria possível construir um léxico aplicável a um domínio específico, como uma determinada área científica, a partir da informação de uma só classe. Esta vantagem é tanto mais importante quanto se tiver em conta a natureza generalista do *corpus*.

Importa, ainda, referir que a escolha de basear este trabalho em Piao *et alii* (2005) se justifica também por, no âmbito do PLN, este tipo de classificação semântica já ter sido testado, como referem os autores, em várias línguas (como no finlandês e no russo) para além do inglês. Assim, pode-se equacionar uma futura comparação de resultados entre as várias línguas. Para além disso, a opção de adaptar o léxico semântico de Lancaster, em detrimento das outras propostas de classificação referidas, deve-se ao facto de o tipo de metodologia usado por Piao *et alii* (2005) ainda não ter sido aplicado ao Português Europeu, ao contrário da metodologia subjacente à WordNet.

Considerando os aspectos anteriormente descritos, a classificação semântica aqui proposta (cf. Quadro 1) é, pois, uma adaptação da de Piao *et alli* (2005), tendo em conta as características do CETEMPúblico. Deste modo, se, por um lado, à semelhança do léxico de Lancaster, a classificação apresenta uma organização em classes e subclasses, por outro, cada classe só admite dois níveis hierárquicos (e não três). Como exemplo, atente-se que, em Piao *et alii* (2005), a classe "Time" tem como uma das suas subclasses "General" e que esta está dividida em "Past", "Present;simultaneous" e "Future", enquanto, na proposta aqui apresentada, as subclasses de "Tempo": "Geral", "Período e "Idade", não admitem subdivisões. Esta última opção relaciona-se com a não produtividade de uma multiplicação de níveis, na medida em que isso levaria a uma excessiva granularidade que não permitiria a pretendida concepção do mundo tão geral quanto possível, ou seja, uma análise de conteúdo. Além disso, assinale-se que a adaptação incidiu também nas designações e códigos atribuídos às classes e subclasses, numa tentativa de melhor adequar a classificação proposta ao *corpus* em causa.

[11] Refira-se que, apesar de terem sido considerados, os traços Humano e Colectivo não estão incluídos na classificação em si, mas apenas na etiquetagem, pois se entende que estes traços devem preceder a notação semântica proposta, na medida em que não são hierarquicamente equivalentes às classes.

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)	

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

4.1.2. Critérios de Anotação

Após a definição dos critérios de adaptação da classificação, e mediante uma primeira observação dos dados, revelou-se necessário estabelecer alguns critérios particulares de anotação, nomeadamente uma classificação não única das ULM. Esta decisão decorre principalmente da verificação da produtividade da polissemia neste material linguístico. Exemplos claros deste aspecto são a inclusão sistemática de uma mesma ULM: (i) nas classes "Organizações, Instituições e Empresas" e "Habitação e Edifícios" ou "Local e Espaço" (Jurafsky & Martin, 2008); e (ii) nas classes "Avaliação e Validação" e "Disposições Emocionais, Atitudes e Comportamentos". A referida classificação não única tem, porém, uma limitação a três classificações, como em Piao *et alii* (2005). Tal limitação justifica-se pela não viabilidade de uma excessiva granularidade, assim como já explicitado a propósito dos níveis hierárquicos.

Refira-se que, perante a necessidade de atribuir mais do que uma classificação, o critério adoptado para determinar aquela que figuraria em primeiro lugar foi o contexto. Com a observação do contexto, pretendia-se, pois, aferir qual das classificações possíveis ocorria com maior frequência. Cabe aqui indicar como excepção a opção de colocar como última classificação as classes "Estatutos, Grupos e Filiação"; "Eventos"; "Habitação e Edifícios" e "Local e Espaço" sempre que estas co-ocorressem com outra(s). Esta opção prende-se com o facto de se pretender uma análise mais eficaz destas sequências num futuro pós-processamento automático.

Ressalve-se que, nos casos em que, no contexto, a frequência de cada classificação candidata era igual ou em que não havia contexto para uma das candidatas, foi necessário fazer uso daquilo que Piao *et alii* (2005) designam *human expert judgement*.

> *For those entries to which multiple candidate semantic categories apply, the categories are arranged in a sorted sequence according to the likelihood and frequency of their application. For each lexeme, usually one or more semantic categories constitute the core or central meaning area while the others form marginal meaning area[s]. In practice, the most likely or common semantic category is put at the front of the candidate list, and the least common one is put at the end of the candidate list. Such a sorting is based on both human expert judgement and empirical statistical information extracted from corpora.* (Piao *et alii*, 2005: 6)

Finalmente, mencione-se que o sistema de notação adoptado consiste num código alfanumérico, tal como se pode verificar na entrada seguinte.

acampamento de refugiados,HumCol:GO3/LO3

No referido código alfanumérico, as duas primeiras letras de cada designação (excepção feita às classes "Comunicação" e "Medidas", para evitar a repetição do código com as classes "Conceitos" e "Metáforas", respectivamente) correspondem às classes

principais e os números (precedidos das letras da respectiva classe) equivalem às subclasses. Quanto aos diacríticos usados, a vírgula separa o lema da(s) classe(s) semântica(s), os dois pontos dividem as designações Humano, Colectivo e Humano Colectivo da notação semântica proposta e a barra oblíqua demarca etiquetas diferentes.

4.2. Aplicação da Classificação

Definidos os pressupostos da classificação e os critérios de anotação, aplicou-se a classificação semântica à lista final de 5068 ULM extraída do CETEMPúblico por meio do tratamento prévio já descrito. A aplicação foi, pois, realizada, por meio de uma etiquetagem manual, à lista de ULM nominais, de modo a testar a abrangência da classificação semântica (isto é, a verificar em que medida a classificação cobre as ULM) e a sua adequação ao tipo de *corpus* seleccionado, ou seja, a validar a taxonomia.

A análise dos resultados desta aplicação deve considerar os seguintes aspectos: (i) abrangência da classificação; (ii) distribuição absoluta das ULM; e (iii) distribuição das ULM por número de classificações.

Quanto à abrangência da classificação, refira-se que a proposta aqui apresentada permitiu classificar 97,1% dos casos (4921 ULM), sendo que os restantes 2,9% (147 ULM) não foram classificados porque implicariam a criação de novas classes não justificadas pela dimensão do *corpus*.

Em relação à distribuição absoluta de ULM por classe e subclasse, observe-se que esta reflecte a natureza do *corpus,* pois os domínios típicos de um jornal generalista têm um maior predomínio. A título de exemplo, as classes "Sociedade", "Governo e Domínio Público" e "Economia" apresentam um maior número de ULM, por oposição às classes "Mundo Animal e Vegetal", "Linguística" e "Astronomia".

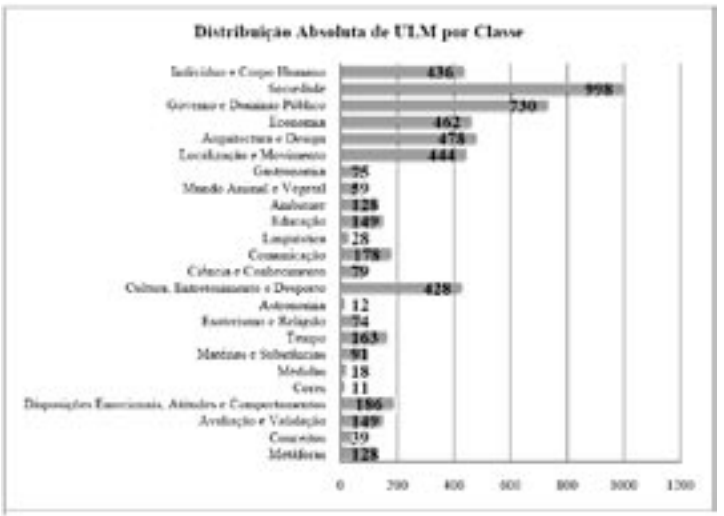

Gráfico 1: Distribuição Absoluta das ULM por Classe

No que respeita à distribuição das ULM por número de classificações, verifique-se que a 86,9% das ULM foi atribuída uma classificação; a 12,8% foram atribuídas duas classificações e apenas a um número residual de 0,3%, três, sendo que estes dados revelam a eficácia da classificação semântica.

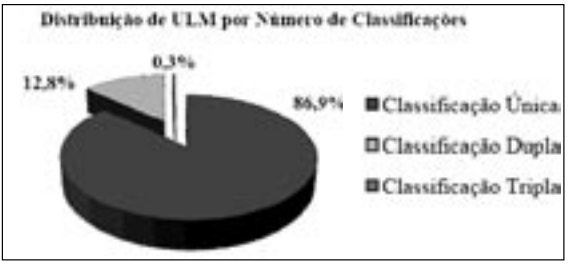

Gráfico 2: Distribuição de ULM por Número de Classificações

Apesar de a maioria das ULM ter sido classificada uma única vez, saliente-se que a terminologia se mostrou clara e flexível no julgamento de casos polissémicos, devido à possível integração de uma mesma ULM em múltiplas classes. Exemplos disso são "agência humanitária" integrada em "Organizações, Instituições e Empresas" e "Habitação e Edifícios" ou "discriminação racial" integrada em "Jurisdição e Justiça" e "Disposições Emocionais, Atitudes e Comportamentos".

Na comparação dos aspectos (ii) e (iii), saliente-se somente que as classes "Localização e Movimento" e "Arquitectura e Design" apresentam um menor número de ULM na distribuição por primeira classificação do que na distribuição absoluta, na medida em que as suas respectivas subclasses "Local e Espaço" e "Habitação e Edifícios" são aplicadas predominantemente como segunda classificação (e não como primeira).

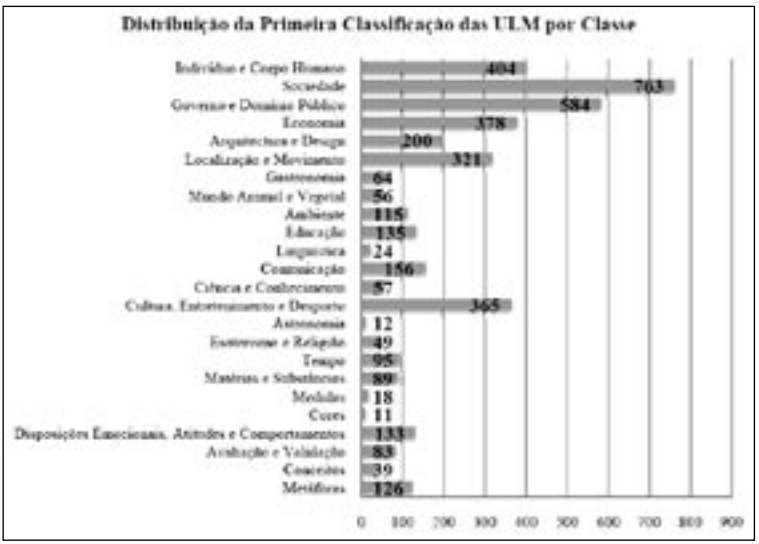

Gráfico 3: Distribuição de ULM por Primeira Classificação

Por último, mencione-se que a validação da classificação pode ainda ser feita pela avaliação da taxa de concordância na anotação de uma amostra por anotadoras diferentes. Assim, assinale-se que uma amostra aleatória de 507 ULM (10% do *corpus*) foi etiquetada separadamente por três anotadoras com experiência linguística (as autoras da classificação semântica) e que os resultados provenientes desse trabalho permitiram validar a classificação semântica, porque se verificou um acordo entre as anotadoras em 95,9% dos casos (das 507 ULM, 486 foram etiquetadas com acordo no que diz respeito às três possíveis classificações em cada palavra).

5. Conclusões

A proposta de classificação semântica de ULM nominais para Português Europeu apresentada pretende ser um contributo para futuros estudos na área da Linguística, mais especificamente da Semântica, bem como do PLN. A necessidade latente de desenvolver léxicos mais completos, ou seja, tratados não só a nível morfossintáctico, mas igualmente semântico, e de extrair e recuperar informação de *corpora* justifica tal expectativa.

O facto de a classificação semântica ter sido aplicada com sucesso a 97,1% da lista de ULM nominais extraída do CETEMPúblico e com uma taxa de concordância entre anotadoras de 95,9% reflecte, pois, a abrangência da classificação aqui proposta. Neste sentido, poder-se-á colocar a hipótese de este trabalho, ainda que exploratório, ser um ponto de partida para futuras avaliações de identificação de ULM e, sobretudo, para trabalhos de classificação semântica de ULM em *corpora* de dimensões e características diferentes.

Finalmente, considera-se que se estabeleceu uma linha de continuidade com trabalhos realizados para outras línguas, que fazem igualmente uso da proposta do léxico semântico de Lancaster (Piao *et alii*, 2005), e se trouxe para a discussão uma proposta diferente das existentes a nível nacional e, por conseguinte, algo inovador em Português Europeu, no âmbito da Linguística e 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 na extracção automática das ULM.

Referências

Baptista, Jorge (1995) *Estabelecimento e Formalização de Classes de Nomes Compostos*. Dissertação de Mestrado, Faculdade de Letras da Universidade de Lisboa.

Bergström, Magnus & Neves Reis (2007) *Prontuário Ortográfico e Guia da Língua Portuguesa*. Lisboa: Casa das Letras (48.ª Edição).

Bick, Eckhard (2006) Noun Sense Tagging: Semantic Prototype Annotation of Portuguese Treebank. In Jan Hajic e Joakim Nivre (eds.) *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*. Praga.

Cunha, Cunha & Lindley Cintra (2005) *Nova Gramática do Português Contemporâneo*. Lisboa: Edições Sá da Costa (18.ª Edição).

Gross, Maurice (1988) Degré de figement des noms composés. *Langages, 90*, Paris: Larousse.

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

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

Mota, Cristina, Paula Carvalho & Elisabete Ranchhod (2004) Multiword Lexical Acquisition and Dictionary Formalization. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, *COLING 2004*. Genebra.

Mota, Cristina, Diana Santos & Elisabete Ranchhod (2007) Avaliação de reconhecimento de entidades mencionadas: princípio de HAREM. In Diana Santos (ed.) *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa: IST-Press.

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

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

Pianta, Emanuele, Luisa Bentivogli & Christian Girardi (2002) MultiWordNet: developing na aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore.

Ranchhod, Elisabete (org.) (2001) *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações. Lisboa: Editorial Caminho.*

Ranchhod, Elisabete (2003) O Lugar das Expressões 'Fixas' na Gramática do Português. In Ivo Castro & Inês Duarte (orgs.) *Razões e Emoção. Miscelânea de estudos oferecida a Maria Helena Mira Mateus*. Lisboa: Imprensa Nacional-Casa da Moeda.

Ranchhod, Elisabete & Paula Carvalho (2003) Unidades Lexicais Complexas. Problemas

de Análise e Etiquetagem. In *Actas do VIII Simpósio International de Comuicación Social*. Santiago de Cuba.

Ranchhod, Elisabete & Cristina Mota (2007) Unidades Lexicais Multipalavra, um osso duro de roer. In Diana Santos (org.) *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa: IST-Press

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

Santos, Diana (2001) Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse.

Sinclair, John (1991) *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Villalva, Alina (2003) Formação de Palavras: Composição. In Maria Helena Mira Mateus *et alii*. *Gramática da Língua Portuguesa*. Lisboa: Editorial Caminho (5.ª Edição).