Desenho de um corpus de especialidade: a propósito do projecto *TermiNáutica*

Margarita Correia (FLUL[1] / ILTEC / SILEX – UMR CNRS)
&
Ana Rebello de Andrade (ILTEC / FCT[2])[3]

1. Introdução

A linguística de corpora é um fenómeno relativamente recente na história da linguística, tendo-se desenvolvido de forma inequívoca com o advento da nova era de micro-informática, através do aumento de capacidade e velocidade no armazenamento / processamento dos dados e da criação de redes informáticas (Cf. Habert, Nazarenko & Salem 1997: 143).

A linguística de corpora, que tende a assumir-se como uma disciplina da ciência linguística, revolucionou, também, a forma de «fazer linguística», na medida em que trabalhar com um corpus linguístico significa analisar uma fonte de dados real e rica da língua, sem hipóteses teóricas sofisticadas, *a priori*, para *a posteriori* formular caminhos teóricos.

Na prática, trabalhar com corpora significa poder dispor de dados linguísticos atestados, ou seja «ter à mão» uma dimensão real da língua em uso de valor individual e social. (Cf. Neto 1995: 27)

O termo 'corpus linguístico' será usado para designar um conjunto finito de textos escritos e orais, que foi seleccionado, arquivado, classificado e processado por computador com o objectivo de tratar um determinado facto linguístico (Cf. Rebello de Andrade 1995: 9)

O projecto *TermiNáutica* é um projecto subsidiado pela Fundação para a Ciência e a Tecnologia e pelo Instituto Camões, através do Programa Lusitânia (Ref.ª PLUS /

[1] Faculdade de Letras da Universidade de Lisboa.
[2] Fundação para a Ciência e a Tecnologia – Bolseira de Doutoramento.
[3] A participação neste Simpósio foi apenas possível graças à atribuição de duas bolsas às autoras, pelo Instituto Camões (www.instituto-camoes.pt), ao abrigo do Programa Lusitânia.

1999 / LIN / 15155). Teve início em Abril de 2001 e os seu termo está previsto para o final de Março de 2003). Da equipa do projecto fazem parte, para além das autoras, Susana Mesquita de Deus Correia, bolseira de investigação.

O projecto *TermiNáutica* tem como principal objectivo a construção de um corpus de especialidade das Ciências Náuticas, em português europeu, etiquetado em SGML, de acordo com as directivas da *Text Encoding Initiative*, com cerca de 1,5 milhões de ocorrências. Nasceu da necessidade efectiva de se constituir um conjunto de dados linguísticos numa área específica - a náutica - e lacunar relativamente ao português europeu. A partir da constituição deste corpus, e dele decorrente, poderão surgir estudos de carácter teórico ou aplicado, tais como, por exemplo, a organização da estrutura conceptual de parte ou da totalidade do vocabulário da náutica[4], estudos sobre variação sociolectal e/ou tecnolectal, estudos diacrónicos (evolução do vocabulário em estudo), entre outros.[5] Do ponto de vista dos estudos aplicados este corpus poderá servir de base para a construção de corpora paralelos[6] – o que poderia beneficiar trabalhos de tradução, de linguística comparada, etc. Este corpus poderá, também, servir para a elaboração de produtos lexicográficos especializados, tanto a nível da escolha (com base em critérios de frequência, pertinência ou outros) dos termos a incluir, como a nível dos contextos definitórios e dos exemplos a utilizar nessas ferramentas linguísticas.

A escolha do domínio das Ciências Náuticas para a constituição de um corpus de especialidade não é inocente. Por um lado, Portugal foi um dos países que, a nível mundial, mais contribuiu para o desenvolvimento destes domínios de saber, o que leva a crer que a sua terminologia será predominantemente constituída por termos vernáculos ou de importação muito antiga, ao contrário do que acontece com outros domínios de experiência, de história mais recente, e desenvolvidos em Portugal em consequência de importação de ciência e de tecnologia. Ao conceber o *TermiNáutica*, pretendeu-se, portanto, contribuir para um conhecimento mais aprofundado do vocabulário de um domínio muito caro à história e à cultura portuguesas, que não tem sido alvo da desejável atenção da parte dos linguistas.

[4] Um dos objectivos visados no âmbito da investigação de doutoramento de Ana Rebello de Andrade (em curso), a apresentar à Universidade de Lisboa.
[5] A equipa pretende desenvolver outros projectos na mesma linha, no âmbito de programas específicos de investigação e desenvolvimento.
[6] Por 'corpora paralelos' entende-se corpora bilingues cujas dimensões, variação, fontes etc., sejam equivalentes em duas (ou mais) línguas.

Nesta comunicação pretendemos dar conta, por um lado, dos pressupostos teórico-metodológicos assumidos no projecto *TermiNáutica* e, por outro lado, do desenho do corpus *TermiNáutica*, procurando justificar as opções feitas.

2. Pressupostos teórico-metodológicos

A teoria lexical em geral e a teoria terminológica em particular têm conhecido, nos últimos anos, significativo desenvolvimento, sendo hoje em dia (re)discutidos princípios que até agora eram praticamente tidos como verdades indiscutíveis. A título de exemplo, refira-se a discussão em torno do estatuto e da caracterização dos termos científicos e técnicos (ou unidades lexicais especializadas) no seio da componente lexical da língua e no âmbito da ciência linguística – Cf. Cabré & Adelstein 2001, Díaz Rojo 2001, Temmermann 2000).

Parece relevante, portanto, esclarecer alguns dos pressupostos teórico--metodológicos que enformam o trabalho desenvolvido no âmbito do projecto *TermiNáutica*, que podem ser resumidos nos seguintes tópicos:

a) As unidades terminológicas (ou "unidades lexicais especializadas") são unidades lexicais *de facto*, merecendo, por isso, ser alvo de análises diferenciadas, que podem ser levadas a cabo também da perspectiva da linguística (a par de outras perspectivas possíveis, tais como a ontologia, a sociologia da linguagem ou as ciências cognitivas);

b) Como unidades lexicais *de facto*, as unidades lexicais especializadas são alvo de variação quer cronológica, quer sociológica, quer, ainda, geográfica, e essa variação deve ser alvo de descrição linguística;

c) Sendo as unidades lexicais especializadas unidades lexicais *de facto*, entende-se que a Terminologia, na sua vertente teórica, se insere na Lexicologia e, na sua vertente aplicada, a Terminografia, se insere na Lexicografia;

d) As unidades lexicais especializadas devem ser estudadas no seu uso (e não em abstracto), o que implica que o desenvolvimento de estudos neste domínio obriga forçosamente à construção de corpora de especialidade devidamente desenhados;

e) O trabalho em terminologia não se limita, portanto, a uma mera tarefa de normalização, tendo uma vertente descritiva fundamental para o desenvolvimento da própria ciência linguística e para a promoção geral do conhecimento humano.

Os princípios anteriormente enunciados distanciam-se da chamada Teoria Geral da Terminologia (Cf. Fedor de Diego, 1995, Arntz & Picht 1996, Cabré dir. 1996), aproximando os trabalhos a desenvolver com base no *TermiNáutica* das abordagens comunicativa (TCT – Cf. Cabré 1999, Cabré & Feliu eds. 2001) e sociocognitiva da Terminologia (Cf. Temmermann 2000).

3. Desenho do corpus

O *TermiNáutica* é um **corpus de referência**, cujos textos são, portanto, amostras. Porém, não foi delimitada uma dimensão pré-determinada para as amostras recolhidas e, sempre que possível, são inseridos no corpus os textos de obras integrais (exs.: manuais de navegação ou manuais de observação meteorológica). A opção por um corpus de referência prendeu-se com o facto de a equipa pretender obter, no final, uma linguagem de especialidade representativa e, por isso, percentualmente seleccionada. Porém, o facto de não termos definido uma dimensão pre-determinada para os textos tem a ver com a diversidade dos textos seleccionados: uns bastantes curtos (exs.: entradas de dicionários de especialidade e de enciclopédias) a par de outros bastante longos (exs.: os referidos manuais, com uma centena ou mais de páginas). Além disso, entendeu-se que, tratando-se de um **corpus de especialidade**, seria muito arriscado seleccionar apenas excertos de obras, dado que poderíamos incorrer no erro de rejeitar segmentos de textos eventualmente mais ricos em terminologia e/ou em características discursivas próprias da especialidade.

O *TermiNáutica* é também um **corpus aberto**, pois embora a sua dimensão esteja, *a priori*, determinada (cerca de um milhão e meio de ocorrências), ela pode vir a sofrer alterações, a nível de incremento de número de ocorrências ou de modificação dos textos nele incluídos, em função das necessidades que a cada momento se forem afigurando.

As opções relativas ao seu desenho articulam-se em torno dos seguintes **critérios de selecção**:

1. dimensão e variação;
2. tipologia dos textos a incluir;
3. selecção dos autores;
4. datações dos textos;
5. inclusão de traduções;
6. inclusão de discurso oral - formal.

3.2.1. Dimensão e variação

A dimensão do corpus *TermiNáutica* fundamentou-se na prática de constituição de corpora de especialidade, onde a dimensão de 1,5 de ocorrências parece ser uma baliza aceitável quando se trata da constituição «nuclear» de um corpus de língua especializada. O facto de se considerar este corpus um «núcleo» para posteriores acrescentamentos, levou-nos a optar pela modelo aberto.

A noção de corpus não deve ser desenquadrada da noção de variação como confirmam Leech, Garside & Atwell (1983: 25): "*In fact the question « How large?» is meaningless unless it is combined with the question of what different types of text are represented in the corpus*."

O *TermiNáutica* é, de facto, um corpus variado, uma vez que são incluídos textos de tipos diversos tais como:

a) manuais;
b) outras obras didácticas;
c) dicionários/glossários e enciclopédias;
d) comunicações de especialistas;
e) textos de "cultura geral" sobre o tema (por ex. obras de história dos Descobrimentos);
f) apontamentos de/para aprendizes;
g) programas de disciplinas leccionadas nas escolas da área;
h) direito marítimo (oral e escrito).

A variedade revela "a língua em acção", enquanto que a quantidade aumenta, de forma decisiva, a possibilidade de ocorrência daquelas unidades lexicais/termos menos

frequentes e, no entanto, representativos do registo linguístico que se pretende descrever.

3.2.2. Tipologia dos textos

Os textos seleccionados têm em conta o tipo de linguagem utilizada e articulam-se basicamente entre textos científicos e técnicos. São também tidos em conta os **diferentes níveis discursivos atestados**, indo do nível mais formal (ex.: comunicações a congressos) ao nível de divulgação (ex.: artigos de enciclopédias).

Nesse sentido estabeleceu-se uma tabela de classificação dos textos, com base em Cabré (1999: 85), que refere que todo o processo de comunicação especializada comporta um determinado grau de variação, desdobrando-se em três níveis - máximo, médio e mínimo -. No projecto *TermiNáutica*, adaptou-se a proposta em epígrafe, sendo admitidos três graus de variação:

Grau 1: textos altamente especializados, de convenção internacional, certificações, normas, etc.

Grau 2: textos presentes em manuais, instruções e conferências;

Grau 3: artigos de enciclopédias, história dos Descobrimentos, textos de vulgarização e textos de debate parlamentar.

3.2.3. Selecção de autores

Os autores dos textos que constituem o *TermiNáutica* são seleccionados de acordo com critérios previamente estabelecidos, tais como a importância e/ou representatividade na área, a idade, o sexo, a nacionalidade – sendo dada a **preferência a autores portugueses**. De resto, optou-se por excluir do *TermiNáutica* textos em norma brasileira, o que poderia conduzir a uma dispersão do objecto de estudo, que é fundamentalmente, o português europeu.[7]

3.2.4. Datações dos textos

Os textos usados como fontes são exclusivamente **textos do séc. XX**. Muito embora tivéssemos inicialmente pensado em nos restringirmos a textos da segunda

[7] Justifica-se relembrar que, embora falando a mesma língua, as comunidades científicas e técnicas portuguesa e brasileira vieram a criar / adoptar terminologias por vezes bastante diferenciadas.

metade do século XX, rapidamente verificámos que a maioria das obras escritas neste domínio do saber em língua portuguesa foram editadas em épocas anteriores, o que nos levou a alargar o escopo cronológico do nosso trabalho.[8]

Este alargamento teve consequências ao nível do desenvolvimento do projecto. Por um lado, este alargamento dificultou as tarefas de digitalização e revisão dos textos. Por outro lado, porém, este alargamento trouxe uma consequência positiva não negligenciável, dado que foi possível verificar que a linguagem náutica, e a da navegação, em particular, sofreu uma evolução drástica ao longo do século XX, fruto do desenvolvimento de tecnologias paralelas (comunicações, observação meteorológica, observação de imagens de satélite), levando a perceber que, até do ponto de vista do desenvolvimento da teoria lexical e terminológica, o estudo deste vocabulário pode desempenhar um papel importante.

3.2.5. Inclusão de traduções

Apesar de a intenção inicial ter sido trabalhar com textos originalmente escritos em português, acabámos por recorrer, embora pontualmente, a traduções para o português devido, sobretudo, ao facto de se ter verificado que a bibliografia mais recente nos subdomínios visados é bibliografia originalmente escrita noutras línguas, nomeadamente em língua inglesa. A inclusão de traduções obedece, porém aos seguintes parâmetros:

a) fidedignidade da editora;
b) características do tradutor (formação, especialidade, etc.).

[8] De facto, tal facto é compreensível à luz da história portuguesa. A Marinha Naval Portuguesa foi, ao longo dos tempos, perdendo o seu inquestionável prestígio em termos científicos e técnicos, indubitavelmente por razões financeiras, mas também pela independência das ex-colónias portuguesas, tendo-se tornado , progressivamente, uma consumidora de ciência e de tecnologia náutica, mais do que uma produtora.
Por outro lado, importa referir o declínio da Marinha Mercante Portuguesa, resultante também de razões económicas, mas, ainda, da evolução e democratização do transporte aéreo.
Finalmente, como é sobejamente conhecido, técnicas de navegação e de comunicação em alto-mar, desenvolvidas por portugueses, foram progressivamente substituídas por técnicas e tecnologias importadas de outros países, nomeadamente anglo-saxónicos. Refira-se, a título de exemplo, a navegação por GPS (*Global Position System*), a comunicação por satélite e a World Wide Web.. A importação destas tecnologias implicou a importação de terminologias associadas.

3.2.6. Inclusão do discurso oral-formal

A inclusão de dados do oral num corpus é sempre desejável, mesmo nas linguagens de especialidade, porque sem este registo a descrição linguística carece de completude. Assiste-se, nos dias que correm, a uma crescente importância atribuída ao estudo do oral que decorre, por um lado, da necessidade de observação do mesmo para se fazer uma descrição global da língua e, por outro lado, dos trabalhos sobre motivações sociais e variação onde Labov, Weinrich & Herzog concluíram que a ausência de heterogeneidade estruturada numa língua real seria disfuncional. (Cf. *apud* Bacelar do Nascimento: 1987: 8).

A opção pela inclusão de dados orais de apenas um registo – o oral formal recenseado nas transcrições dos debates parlamentares acerca do tema – no corpus nuclear prende-se, sobretudo, com as limitações temporais e financeiras que o projecto sofreu, em virtude das dificuldades e custos que o tratamento de dados orais comportam. O interesse pelo tipo de discurso referido prende-se também com o facto de se querer constituir um subcorpus de direito marítimo, para futuros estudos terminológicos sobre legislação.

Nesse sentido, Miguel de Oliveira, investigador de pós-doutoramento, no ILTEC, encontra-se já a constituir um corpus de direito comercial marítimo, com base em textos recolhidos na Internet e que constituirá um subcorpus do *TermiNáutica.*

Não se exclui, no entanto, a inserção, no futuro, de subcorpora orais variados (transcrições de gravações de profissionais em acção, etc.) no macro-corpus *TermiNáutica* que se pretende construir. Tal como refere Leech (1991: 11): "*The collection of spoken discourse on the same scale as written text will remains a dream of the future*".

3.2.7. Codificação / etiquetagem

O corpus *TermiNáutica* é codificado em SGML (*Standard Generalized Markup Language)* seguindo as propostas da *Text Encoding Initiative* (cf. o endereço URL http://etext.lib.virginia.edu/TEI.html), com o objectivo de fazer deste corpus um produto reutilizável, não apenas no seio do ILTEC, como por membros da comunidade científica, internos ou externos ao Instituto.

4.Conclusões

Com esta comunicação, pretendeu-se, acima de tudo, justificar opções feitas na construção do corpus *TermiNáutica,* tendo em vista a possibilidade de discussão dessas opções. Como foi referido anteriormente, a linguística de corpus é uma forma relativamente recente de fazer linguística, pelo que a discussão dos aspectos metodológicos faz, a nosso ver todo o sentido.

5. Bibliografia

Andrade, A. Rebello 1995. *As palavras importadas no léxico da decoração.* Dissertação de Mestrado apresentada a Faculdade de Letras da Universidade de Lisboa (inédita).

Ajmer, K. & B. Altenberg (eds.) 1991. *English Corpus Linguistics*. London / New York: Longman.

Bach, C. R. Saurí, J. Vivaldi & M. T. Cabré 1997. *El corpus de l´IULA : Descripció*. Sèrie Informes 17. Barcelona: UPF, IULA.

Bacelar do Nascimento, M. F. 1992. *Corpus de Referência do Português Contemporâneo*. Apresentação de projecto ao Programa Lusitânia /JNICT: inédito.

Bacelar do Nascimento, M. F., M.ª L. Garcia Marques & M.ª L. Segura da Cruz 1987. *Português Fundamental – Métodos e Documentos*, tomo 1. Lisboa: INIC e CLUL.

Bacelar do Nascimento, M. F. 1987. *Contribuição para um dicionário de verbos do português*. Lisboa: INIC e CLUL.

Biber, D., S. Conrad & R. Reppen 1998. *Corpus Linguistics – Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Cabré, M. T. 1998. *Terminology, Theory, methods and applications*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Cabré, M. T. 1999. *La Terminología – Representación y comunicación – Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: IULA.

Cabré, M. T. & J. Feliu (eds.) 2001. *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica*. Barcelona: IULA.

Garside, R., G. Leech & T. McEnery (eds.) 1997. *Corpus Annotation*. London / New York: Longman.

Habert, B., A. Nazarenko & A. Salem 1997. *Les linguistiques de corpus*. Paris: Armand Colin.

Kennedy, G. 1998. *An introduction to Corpus Linguistics.* London / New York: Longman.

Leech, G. R. Garside & E. Atwell 1983. *Recent Developments in the use of computer corpora in English language research.* Transactions of the philological society. Oxford : Basil Blackwell

Leech, G. 1991. «The state of the art in corpus linguistics». *In: English Corpus Linguistics*, London and New York: Longman.

Neto; P. 1995. *Combinatórias Lexicais no Discurso da Astronomia – Um Estudo em Estatística Lexical*. Dissertação de Mestrado apresentada a Faculdade de Letras da Universidade de Lisboa (inédita).

Pearson, J. 1984. *Terms in Context.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Sager, J. C. 1993. *Curso práctico sobre el processamiento de la terminología*. Madrid: Fundación Germán Sánchez Ruipérez – Ediciones Pirámide.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amesterdam / Philadelphia: John Benjamins Publishing Company.