O lugar do *corpus* na investigação linguística

Maria Fernanda Bacelar do Nascimento
Centro de Linguística da Universidade de Lisboa

Apesar de existir uma longa tradição de estudos linguísticos realizados com base em *corpora* (primeiro, *corpora* constituídos manualmente, e, nas últimas décadas, *corpora* electrónicos, alguns deles de grandes dimensões) não se pode dizer que o tema desta mesa-redonda seja um tema gasto ou sem futuro; na realidade, só a partir dos anos 90, graças a um extraordinário desenvolvimento dos meios tecnológicos, se começou, verdadeiramente, a generalizar, na comunidade linguística, o reconhecimento das grandes potencialidades destes Recursos Linguísticos para um melhor conhecimento das línguas.

Os *corpora* são, hoje, encarados sob perspectivas epistemológicas diversas que correspondem, no essencial, à seguinte interrogação: os *corpora* constituem, unicamente, um suporte poderoso para a aplicação de novas e mais eficazes metodologias ou, muito mais do que isso, consubstanciam eles, actualmente, um novo ramo da Linguística?

Uns encaram os estudos sobre *corpora* como uma nova abordagem filosófica, uma nova maneira de pensar a língua (cfr., por ex., LEECH, 92), outros admitem já a sua compatibilidade com modelos cognitivistas (cfr., por ex., SHÖNEFELD, 99), outros, ainda (cfr. SINCLAIR, 91) pensam que a evidência dos factos atestados, estimulando novas descrições e hipóteses teóricas, contribuirá decisivamente para que a Linguística venha a atingir um maior grau de maturidade.

Pela nossa parte, consideramos que os *corpora* favorecem essencialmente uma Linguística descritiva, fortemente apoiada pelas novas tecnologias, e permitem tomar como ponto de partida da descrição a análise de quantidades significativas de dados autênticos, à semelhança do que se faz noutros domínios científicos. O uso de *corpora* permite a realização de descrições linguísticas de base empírica e promove, com isso, a discussão de questões teóricas solidamente fundamentadas.

A nosso ver, os *corpora* não constituem, em si próprios, um novo ramo da linguística pelo que não reconhecem a existência de qualquer incompatibilidade entre estudos de *corpora* e estudos teóricos nem entre os métodos indutivo e dedutivo que caracterizam cada um destes estudos. Consideramos, antes, de uma forma abrangente, que os *corpora* proporcionam novas maneiras de estudar as línguas, das quais resultam descrições, generalizações e hipóteses teóricas de grande consistência porque fortemente enraizadas nos dados empíricos. Como diz Dominique WILLEMS, "La question se pose de savoir si l'utilisation de 'nouvelles' données, en particulier des données de corpus, mène à une autre linguistique. Par notre part,

Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2002, pp. 601-605.

la réponse serait nuancée: si les données de corpus constituent un enrichissement considérable du matériau linguistique et permettent d'ajouter des dimensions supplémentaires à l'analyse (telle la dimension statistique et co(n)textuelle), elle ne change pas fondamentalement ni les questions posées, ni les méthodes: si l'objet est partiellement différent, les méthodes rejoignent celles de la linguistique descriptive." (WILLEMS, 2000).

Por estas razões, achamos que um *corpus* se define não só por factores tão importantes como a sua dimensão, constituição, diversificação, estrutura e dinâmica de actualização, mas também, decididamente, pela variedade de utilizações que proporciona. Assim, se o primeiro objectivo do grupo de Linguística de *Corpus*, do Centro de Linguística da Universidade de Lisboa – CLUL, foi constituir um grande *Corpus* de Referência do Português Contemporâneo – CRPC (www.clul.ul.pt/sectores/projecto_crpc.html), os objectivos que se lhe seguiram repartem-se entre divulgação e análise dos dados, sem, no entanto, descurar o desenvolvimento e a actualização dos seus sub-*corpora* oral e escrito.

A descrição dos projectos nacionais e internacionais concluídos ou em curso, com a participação do CLUL e tendo por base o CRPC encontra-se na página do CLUL – www.clul.ul.pt/sectores/projectos.html#2.

Apresenta-se, seguidamente, uma tabela de Recursos Linguísticos – *corpora* e léxicos – em que participou o grupo de Linguística de *Corpus* do CLUL, já disponíveis ou a disponibilizar durante o ano de 2003.

TABELA DE RECURSOS LINGUÍSTICOS DO CLUL JÁ DISPONÍVEIS

	Tipo de acesso	Pequena descrição		Financiamento e Autoria	Disponível em
Corpora	**Consulta on-line**	Sub-*corpus* do CRPC (4.646.737 palavras)	*Corpus* ELAN (2. 989. 746 palavras) – Jornal, Revista, Livro técnico, Livro Literário, Varia – Concordâncias e frequências.	Programa MLIS 121: Parceria europeia (Portugal: CLUL)	www.clul.ul.pt/sectores/projecto_rld1.html
			Sub-*corpus* do Projecto Recursos Linguísticos (1.656.991 palavras) – Livro literário – Concordâncias e frequências.	FCT-Programa Lusitânia e FCG: CLUL e SPA	www.clul.ul.pt/sectores/projecto_rld1.html
	Download	*Corpus* PF Publicado (106.488 palavras) – Oral – Acesso a todo o *corpus*		CLUL	www.clul.ul.pt/sectores/corpus_oral_pf_publicado.zip
	CD-ROM	*Português Falado – Documentos Autênticos* (86 conversas, 8h 44m, 91.966 palavras) – Oral – Acesso ao *corpus*: som e transcrição alinhados.		Programa LINGUA/SOCRATES: CLUL e Instituto Camões	4 CD-ROM editados por CLUL/Instituto Camões
	CONSULTA VIA CLUL	Sub-*corpora* do CRPC (201. 487. 845 palavras) – Oral e Escrito – Concordâncias, frequências e combinatórias		FCG: CLUL	Mediante pedido dirigido ao Coordenador do Projecto
	AQUISIÇÃO	Sub-*corpus* PAROLE (3.000.000 palavras com 250.000 palavras anotadas morfossintacticamente)		CE: Telematics Application Programme: Parceria europeia (Portugal: CLUL e INESC)	www.elda.fr/cata/text/W0024.html

TABELA DE RECURSOS LINGUÍSTICOS DO CLUL JÁ DISPONÍVEIS

Tipo de acesso	Recurso disponível	Financiamento e Autoria	Disponível em:
Consulta on-line e download	LMCP (26. 443 lemas e 140.315 formas) – Léxico do português contemporâneo com informação morfológica e de frequência.	FCT-Programa PRAXIS XXI: CLUL, INESC e Ed. Verbo	www.clul.ul.pt/sectores/lmcpc
	SIMPLE (300 unidades) – Sub-léxico PAROLE com descrição semântica em formato .sgml	CE: Telematics Application Programme: Parceria europeia (Portugal: CLUL e INESC)	www.ub.es/gilcub/SIMPLE/simple.html
Aquisição	PAROLE (20.000 unidades) – Léxico com anotação morfossintáctica e descrição sintáctica.	CE: Telematics Application Programme: Parceria europeia (Portugal: CLUL e INESC)	www.elda.fr/cata/text/L0035.html

TABELA DE RECURSOS LINGUÍSTICOS DO CLUL A DISPONIBILIZAR EM 2003

Recurso Tecnológico	Financiamento e Autoria
Corpus Compartilhado VARPORT (252. 300 palavras) – Oral e Escrito – Português do Brasil e Português Europeu dos sécs. XIX e XX (acesso directo).	ICCTI-Portugal e CAPES-Brasil: CLUL e Universidade Federal do Rio de Janeiro
Corpus do Projecto Recursos Linguísticos para o Português (9. 000.000 palavras) – com 500.000 palavras anotadas morfossintacticamente (consulta).	FCT – Programa Lusitânia e FCG: CLUL e SPA
Corpus C-ORAL-ROM – *Corpus* comparável multilingue – 300.000 palavras de cada uma das 4 línguas: Espanhol, Francês, Italiano, Português – Oral – Discurso formal e informal (CR-ROM).	CE – IST Programme: Parceria europeia (Portugal: CLUL)
REDIP – *Corpus* de PortuguêsEuropeu recolhido nos media: gravações de rádio e televisão e amostragens de jornais (consulta).	FCT – Programa Lusitânia; ILTEC Universidade Aberta e CLUL

Bibliografia

LEECH, G. (1992) "Corpora and theories of linguistic performance", *Directions in Corpus Linguistics*, ed. by Jan Svartvic, Berlin, Mouton de Gruyter, pp.105–122.

SHÖNEFELD, D. (1999) "*Corpora* Linguistics and Cognition", *International Journal of Corpus Linguistics*, Amsterdam/Philadelphia, John Benjamins Publishing Company, vol. 4–1, pp. 137–171.

SINCLAIR, J. (1991) " *Corpus, Concordance, Collocation*", Oxford, Oxford University Press.

WILLEM, D. (2002) "Objet d'étude, théories et données sur la place des corpus dans la recherche linguistique contemporaine", BILGER, M. (ed.), *Corpus, Méthodologie et applications linguistiques*, Paris, Honoré Champion, pp. 149–155.