Constituição de *corpora* de especialidade

M. Rute V. Costa
Universidade Nova de Lisboa
Faculdade de Ciencias Sociais e Humanas

O *corpus*, objecto de estudo que está na origem das linguísticas de *corpora*, é um lugar de observação que permite a descrição de actualizações da língua organizadas em enunciados, discursos ou textos. Na base da constituição destes conjuntos de dados linguísticos estão critérios de selecção sistematizados, que facultam a legítima atribuição do estatuto de *corpus* a tais conjuntos de dados.

O facto de o termo "*corpus linguistics*" não ser utilizado nos textos anteriores a Chomsky não significa que os *corpora* não fossem usados e explorados com a finalidade de análise linguística. Assim, recorrer aos *corpora* como objecto de análise não é um procedimento inovador. Em 1951, Harris considerava já o *corpus* o único objecto legítimo da linguística e designava por linguística estrutural a investigação que operava, *a priori* ou *a posteriori*, com *corpora*.

Aarts, por sua vez, considera que o conceito de "*corpus linguistics*" não dá conta de uma actividade totalmente nova em linguística: "[...] *if we take corpus linguistics as referring to linguistic research based on observed utterances, we can say that this type of research has a very long history indeed. Only in earlier days it was simply called linguistics of philology*" (Aarts, 1990:13).

Chomsky (1957, 1965) modificou o objecto da linguística, considerando que os *corpora* não poderiam nunca ser entendidos como objectos de análise úteis para o linguista; privilegia, recorrendo à introspecção, uma aproximação racionalista ao objecto, em detrimento de uma aproximação empírica: "*Chomsky changed the object of linguistic enquiry from abstract descriptions of language to theories which reflected a psychological reality, cognitively plausible models of language*" (McEnery, Wilson, 1997: 4). Para este autor, os *corpora* dão conta exclusivamente dos actos de *performance*, não revelando os actos de competência, que podem ser unicamente determinados pelo falante. No que concerne ao *corpus*, McEnery e Wilson sintetizam o posicionamento de Chomsky do seguinte modo: "*A corpus is by its very nature a collection of externalised utterances; it is performance data, and such it must of necessity be a poor guide to modelling linguistic competence*" (1997:5).

Nos anos sessenta, as metodologias e as teorias associadas aos *corpora* ganham uma nova dinâmica. Vários autores (Aarts, 1990; Leech, 1997; Habert, 1998) consideram esta década um marco na história recente das linguísticas de

***Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2002, pp. 253-258.**

corpora: "*The year of 1961, which more famously saw then first manned space flight, is the date to which corpus linguistics can look back as the date when the entreprise now known as corpus linguistics (or more precisly computer corpus linguistics) came into being*" (Leech, 1997:1).

É a escola anglo-saxónica que populariza o termo "*corpus linguistics*", que recobre simultaneamente o objecto, bem como as metodologias e as teorias que se constroem a partir dos *corpora*. A própria definição de *corpus* é actualizada tendo por referência o suporte em que este é armazenado, o seu formato electrónico, incrementando as perspectivas de análise: "*In the past thirty-five years, the term corpus has been increasingly applied to a body of language material which exists in electronic form, and which may be processed by computer for various purposes such as linguistic research and language engineering*" (Leech, 1997:1).

Também Sinclair (1996) propõe uma definição de "*computer corpus*", independente da de *corpus*: "*A computer corpus is a corpus which is encoded in a standardised and homogenous way for openended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance*" (Sinclair, 1996:6). Tal definição pressupõe, implicitamente, que o *corpus* seja entendido como: "[...] *collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language*" (Sinclair, 1996:6).

O *corpus* informatizado pode apresentar-se sob duas formas, isto é, na sua forma bruta ("*raw corpus*") ou anotada. Enquanto que o *corpus* bruto é um objecto para testar hipóteses "[...] *the test bed for his hypotheses about the structure of the language, which he has expressed in a formal grammar*" (Aarts, 1990:18), o *corpus* anotado é enriquecido com informação de diversa natureza: morfológica, sintáctica, semântica, prosódica, crítica, etc., e "[...] *serves as a linguistic database for all linguists studying the structure of the language*,[...]" (Aarts, 1990:18).

Pensamos, deste modo, que as linguísticas de *corpora* assumem um duplo estatuto: por um lado, de sub-disciplina no seio da linguística, por outro, funcionam como disciplina auxiliar para todas as restantes disciplinas da linguística: "*It creates textual databases which have been enriched with detailed morphological and syntactical information and, where possible, with phonological and semantic information. Within the foreseeable future, every linguist will be able to make use of such databases, which means he will also have at immediate numerical data about the use of constructions and sentence patterns, the realisation of grammatical sentences, etc.*" (Aarts, 1990:16).

Para que os resultados obtidos a partir de *corpora* sejam fiáveis, é indispensável que o objecto sobre o qual recaem as nossas hipóteses seja adequadamente definido e delimitado. Com o aumento crescente e variado dos *corpora*, surge a necessidade de reflectirmos adequadamente, por um lado, sobre as características do *corpus* merecedor dessa designação, por outro saber como classificar a diversidade resultante de tal proliferação.

Não devemos considerar todo e qualquer tipo de *corpora* um objecto válido, *a priori*, para todos os fins da análise linguística. Os critérios de selecção dos enunciados que compõem o *corpus*, assim como as suas propriedades, têm de estar, necessariamente, em consonância com os objectivos pré-estabelecidos pelo linguista.

Os *corpora*, tal como os enunciados que os constituem, podem ser classificados em tipos distintos. Aludiremos, exclusivamente, à tipologia de *corpora* de textos escritos.

Em 1996, nas "*Preliminary recommendations on Corpus Typology*", Sinclair propõe uma tipologia de *corpus*, que entendemos estar arquitectada sobre pressupostos metodológicos e teóricos que determinam o seu desenho e a sua constituição. Um *corpus* deve ser edificado tendo por base parâmetros que permitam legitimar os resultados obtidos a partir da sua análise. Subjacentes à compilação de um *corpus*, estão respostas a perguntas que Kennedy equaciona do seguinte modo: "*Issues have included whether a corpus should be a static or dynamic sample of a language, how best it can be representative of a language or a genre, how big a corpus should be to be representative or to serve particular purposes, and how big the text samples should be*" (Kennedy, 1998:60).

O facto de os *corpora* poderem ser constituídos por extractos de textos ou por textos integrais, leva Sinclair (1996) a evitar a utilização do termo texto em favor da expressão "*pieces of language*", opção terminológica que indicia, no que concerne ao conceito de texto, um posicionamento teórico não explícito, "*Note that the non-comminttal word 'pieces of language' is used above, and not 'texts'. This is because of the question of sampling techniques used. If samples are to be all same size, then they cannot all be texts. Most of them will be fragments of texts, arbitrarily detached from their contents*" (Sinclair, 1996:6). Com esta observação, Sinclair faz uma tímida incursão nas questões teóricas relativas ao texto, admitindo que os *corpora* podem ser constituídos por outras realidades que não textos, nunca chegando a explicitar o seu entendimento do mesmo. É ainda com alguma surpresa que verificamos que Sinclair e Ball nem mesmo no documento "*Preliminary recommandations on Text Typology*" (1996) definem o conceito de texto, sendo a supracitada asserção desprovida de valor, uma vez que não revela um posicionamento claro face às teorias do texto.

Convictos da indispensabilidade de assumirmos um posicionamento teórico inequívoco no que se refere ao conceito de texto (cf. ponto 2.2.1), e por entendermos não corresponder a expressão "pieces of language" de Sinclair (1996) a um conceito da linguística, recorremos, neste momento, para mencionar os dados linguísticos que integram os *corpora*, ao conceito de enunciado na acepção de Greimas, isto é, "[...] *toute grandeur pourvue de sens, relevant de la chaîne parlée ou du texte écrit, antérieurement à toute analyse linguistique ou logique*" (Greimas, 1979:123).

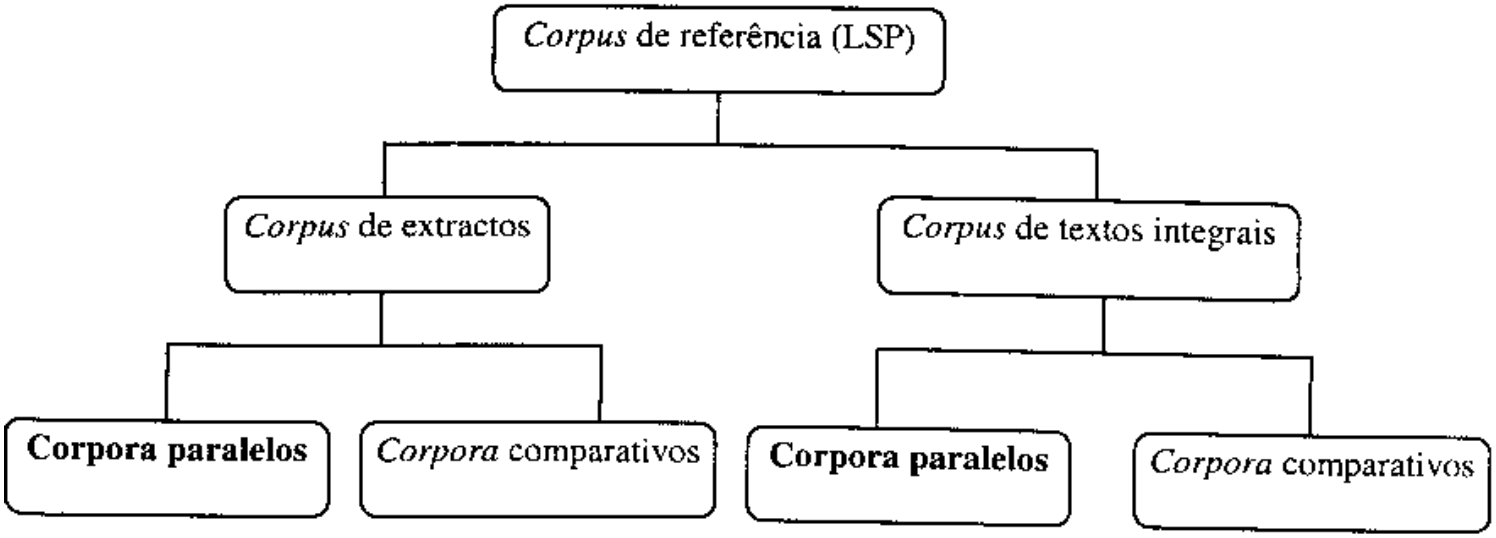
Deliberadamente, excluímos da tipologia que apresentamos os *corpora* lexicográficos, isto é, os *corpora* de frequências, os *corpora* de palavras, etc...

Sinclair utiliza os termos "*sublanguage*" e "*special corpora*" para designar todos os enunciados que são produzidos em situações sociais específicas, devendo, além dos critérios internos, recorrer-se também a critérios externos para definir tais conceitos que se opõem, pela sua natureza, ao de *corpus* ao de referência: "*A sublanguage is at the other end of the linguistic spectrum from a reference corpus*" (Sinclair, 1996:20). Assim, deduzimos que todo o enunciado que, devido às suas características, não pode integrar o *corpus* de referência, faz parte integrante de *corpora* especiais. Baseando-se em Sinclair, Pearson apresenta-nos alguns exemplos de *corpora* especiais: "*Examples of special corpora given by Sinclair are corpora of the language of children, the language of geriatrics, the language of non native-speakers and the language of very specialized areas of communications*" (Pearson, 1998:46). Geralmente, quando tais conjuntos de dados são abordados do ponto de vista linguístico, cria-se a expectativa de encontrar desvios à norma. É a constatação desses desvios que impede de os incluir no *corpus* de referência, em razão de o resultado da extracção de um enunciado deste tipo não constituir um *subcorpus*, porque: "*A subcorpus has all the properties of a corpus but happens to be part of a larger corpus*" (Sinclair, 1996:9).

Não podemos, no entanto, aceitar que enunciados resultantes de situações de comunicação entre especialistas sejam incluídos num *corpus* especial, ou sejam entendidos como uma sublíngua, já que de forma alguma podemos considerar que nestes enunciados existem desvios à norma. O que encontramos neste tipo de enunciados são termos utilizados com acepções diferentes e construções morfossintácticas que, idealmente, reflectem comunicações monorreferenciais e não ambíguas, inerentes às situações de comunicação especializada: "*Les corpus spécialisés réunissent des données linguistiques relatives à une dimension particulière: un domaine, un thème, une situation de communication*" (Habert et alii, 1998:37)

Consideramos que os discursos proferidos em situações de especialidade constituem *corpora* de especialidade, que requerem aproximações metodológicas e teóricas particulares.

Assim, o conjunto dos enunciados produzidos por especialistas em Detecção Remota constituem um *corpus* de especialidade. Se o conjunto de enunciados de especialidade for representativo dos enunciados produzidos pela classe profissional em causa e se a quantidade de enunciados recolhidos for significativa, então assumimos estar perante um *corpus* de referência de especialidade:



A representatividade dos enunciados que compõem um *corpus* é uma questão central da constituição dos *corpora* de especialidade, uma vez que a diversidade dos textos no seio de uma área de especialidade é imensa: "[...] *l'étroitesse du sujet n'empêche pas sa diversité interne, et le problème ne consiste là encore à concevoir un corpus équilibré, que l'on puisse considérer comme échantillon de l'ensemble que l'on veut analyser*" (Habert et alii, 1998:37).

A noção de representatividade em *corpora* especializados não pressupõe a noção de quantidade, dado que a produção de textos numa área de especialidade, numa língua determinada, pode ser diminuta, assumindo o tamanho do *corpus* um valor relativo.

Possuir uma colecção de textos informatizados não é condição suficiente para que possamos considerar estar em presença de uma corpus; para o constituir é necessário ter em conta um conjunto de pressupostos teóricos e metodológicos considerados de importância fundamental.

Bibliografia

Aaarts, Jan (1990), "Corpus Linguistuics: An Appraisal", *Computers in Letrary and Linguistics Research*, Proceedings of the International Conference of the Association for Literary and Linguistic Comptuing, Paris-Genève, Hamesse, J., Zampolli, A., pp. 13-28.

Greimas, Algirdes Julien; Courtès, Joseph (1979), *Sémiotique. Dictionnaire raisonné de la théorie du langage*, Tome 1, Paris, Hachette, 443 p.

Habert, Benoît; Fabre, Cécile; Issac, Fabrice (1998), *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, Paris, Masson, 320.

Kennedy, Graeme (1998), *Na introduction to Corpus Linguistics*, Longman, London and New York, 315 p.

Leech, G. (1997), "Grammatical tagging", *Corpus Annotation, Linguistics Information from Computer Text Corpora*, ed. Roger Garside, Geoffrey Leech, Tony McEnery, London & New York, Longman, pp. 19 –33.

McEnery, Tony e Wilson, Andrew (1996), *Corpus Linguistics*, Ed. Tony McEnery and Andrew Wilson, Manchester, Edinburgh University Press, 207 p.

Pearson, Jennifer (1998), *Terms in Context*, Coll. Studies in Coprus Linguistics, Amsterdam / Philadelphia, John Benjamins Publishing Company, 242 p.

Sinclair (1996), *EAGLES: Preliminary Recommandations on Corpus Typology* (EAG – TCWG –CTYP/P), Version of May, pp. 25 disponível em http://www.ilc.pi.cnr.it