Corpora e Estudos Linguísticos

Maria Francisca Xavier
Centro de Linguística da Universidade Nova de Lisboa

A **Linguística de Corpus** tem já uma longa tradição, embora não fosse assim designada inicialmente. Os estudos linguísticos anteriores a Chomsky (1957) eram fundamentalmente baseados em corpora constituídos a partir de textos escritos. Desde sempre, tanto os estudos históricos como aqueles sobre a linguagem das crianças iniciados já no século dezanove[1] exigem, necessariamente, a constituição de corpora.

A constituição de **corpora** foi a base da **Linguística Empírica**, **Linguística Descritiva**, do **Estruturalismo** de Bloomfield (1933) e do **Distribucionalismo** de Harris (1951). Os estruturalistas americanos tiveram o grande mérito de despertar o interesse pelo estudo das realizações orais das línguas ameríndias, que são línguas muito diferentes daquelas que tinham sido mais estudadas até então, as línguas indo-europeias. A descrição das línguas ameríndias, que os antropólogos e linguistas desconheciam por completo, revelou a existência de aspectos particulares das gramáticas das línguas naturais que ainda não tinham sido observados e identificados noutras línguas. A segmentação e o agrupamento de constituintes e a constituição de paradigmas permitiram identificar os morfemas livres, as raizes, e os morfemas presos, os afixos, assim como os sintagmas de línguas que os linguistas não falavam, mas que decidiram estudar e conseguiram descrever. Foi possível identificar tanto morfemas lexicais como funcionais, listar e classificar palavras e tipos de frases, e foi até possível utilizar com sucesso dados de uma dessas línguas em códigos secretos para a comunicação entre elementos das forças militares americanas durante a segunda guerra.

A Linguística de Corpus avançou, então, através de árduo trabalho de campo, recolhendo corpora tão extensos quanto possível, analisando, classificando e quantificando os dados ainda sem o auxílio de computadores nem de programas para tratamento da língua natural.

A Linguística de Corpus actual tem a vantagem de dispor da possibilidade de utilizar **corpora informatizados** e de **ferramentas computacionais** que permitem tanto a extracção, a ordenação e a contextualização fáceis como a quantificação rápida e ampla de dados empíricos úteis para estudos linguísticos diversos, simultaneamente, descritivos e teóricos.

---

[1] McEnery & Wilson (1996) indicam o período compreendido entre c. 1876-1926 como a fase inicial da investigação sobre a linguagem das crianças baseada em diários de pais que continham registos de enunciados dos filhos.

***Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2002, pp. 869-872.**

Mas, porque o uso frequente das palavras em diferentes perspectivas as torna muitas vezes vagas ou ambíguas, será útil reflectir sobre o que se entende em Linguística por:

(i) corpus
(ii) corpora
(iii) corpus textual (oral/escrito)
(iv) corpus de dados linguísticos

Por outro lado, importa reflectir, ainda, sobre se os corpora poderão ser considerados fechados e/ou abertos?

Segundo Tony McEnery & Andrew Wilson (1996, p. 59) "... **um corpus** implica tipicamente **um corpo finito de texto**, seleccionado para ser o mais **representativo** possível de uma variedade particular de uma língua, e, também, que possa ser **armazenado e manipulado utilizando um computador.**"[2]

Parece ser claro que estes autores não contemplam a oposição entre **corpus textual** vs. **corpus de dados linguísticos,** nem consideram que um corpus possa ser aberto. É, contudo, de notar que corpus textual e corpus de dados linguísticos são instrumentos de trabalho distintos e que ao se considerar a possibilidade de um corpus ser aberto, se assume, inequivocamente, que um corpus tem sempre limitaçoes.

Para clarificar esta questão, podemos também interrogarmo-nos sobre se um texto constitui ou não um corpus textual, ou ainda se um corpus textual será um conjunto de textos ou um conjunto de amostras de textos. Ora um texto é apenas um texto e não deve ser confundido com um corpus, pois muito embora possa conter informação relevante e possa também fornecer exemplos úteis para ilustrar aspectos linguísticos particulares não é, seguramente, suficientemente representativo de uma variedade linguística.

Actualmente, parece ser possível considerar que um corpus de dados linguísticos é, então, quer um conjunto de dados extraído de um corpus textual digitalizado, quer um conjunto de dados construído por introspecção ou por eliciação, também digitalizado, e em ambos os casos submetidos à análise automática e/ou humana, à qual se segue, obrigatoriamente, a validação humana.

Deste modo, um corpus é por definição um conjunto finito de textos ou de amostras textuais ou de dados linguísticos, que pretende ser, em qualquer caso, representativo do objecto de estudo, sendo os corpora, por sua vez, constituídos por mais do que um corpus, todos eles construídos para serem representativos do objecto de estudo.

No entanto, fala-se também em corpus aberto. Neste caso, trata-se de um corpus que se pretende alargar incorporando mais textos, amostras ou dados linguísticos de modo a aumentar a sua representatividade, uma vez que o corpus fechado

[2] A tradução e os realces a negrito são meus.

não é, certamente, suficientemente representativo. Assim, em cada fase de alargamento de um corpus, obtem-se novamente um conjunto finito, um corpus que apresenta um determinado grau de representatividade e que é, modernamente, em formato digital e está preparado para ser processado e, eventualmente, marcado, etiquetado e analisado sintactica e semanticamente. Porém, seja qual for a sofisticação das ferramentas informáticas de processamento da língua natural, nada substitui a indispensável análise humana, a única que pode conduzir a generalizações e a explicações linguísticas.

E, embora um corpus nunca seja suficientemente representativo de uma língua, nem de uma variedade ou estado de língua, menos ainda da sua história, este é um instrumento indispensável para os estudos de Linguística Histórica. Um corpus diacrónico textual, entendido como um conjunto de textos informatizados para estudos diversos, tendo em vista a multiplicidade de fenómenos linguísticos, históricos e culturais que poderão vir a ser estudados, tem de ser, necessariamente, um corpus aberto, melhor dizendo, um corpus em construção.

A constituição do CIPM-Corpus Informatizado do Português Medieval, iniciada há dez anos, bem como os estudos de Linguística Histórica que têm vindo a ser realizados com dados dele extraídos têm exigido frequentes momentos de reflexão sobre a representatividade tanto dos diferentes corpora textuais que a partir dele têm sido constituídos para estudos particulares como dos corpora de dados linguísticos atestados neles. Tem sido ainda a verificação da fraca representatividade dos corpora para determinadas investigações, o que obriga a continuar a construção do CIPM[3].

Para ser possível estudar o léxico histórico, o corpus deverá ser cronologicamente representativo e terá de ser, também, tipologica e tematicamente variado, pois pretende-se extrair dele tanto o léxico comum como as terminologias. Para além de listagens de palavras, com ou sem etiquetas categoriais, de expressões estatisticamente relevantes extraídas de corpora extensos, de concordancias e de quantificacoes e estatísticas, que constituem instrumentos extremamente úteis, o trabalho principal reside na concepção e na definição dos critérios para a elaboração dos diferentes módulos que irão sendo desenvolvidos com o objectivo de realizar um Dicionário do Português Medieval – DPM[4].

Para o estudo de um aspecto particular do léxico, da morfologia ou da sintaxe, por exemplo, deverá ser constituído um subcorpus de um corpus extenso, que é um conjunto de textos ou de amostras de textos criteriosamente seleccionados. Este subcorpus deverá permitir a constituição de um corpus de dados relevantes para a análise e a explicação do aspecto particular que se pretende estudar. Um subcorpus de dados linguísticos será relativamente a um aspecto do léxico um conjunto determinado de ocorrencias lexicais, contextualizadas em concordâncias, e será relati-

---

[3] Cf. Xavier, Crispim e Vicente "Português Antigo – Construção e Disponibilização de Recursos em Suporte Informático" nestas *Actas*.

[4] Cf. Xavier, Vicente e Crispim *orgs.* (1999) e (2002).

vamente a um aspecto da sintaxe um conjunto de frases seleccionado a partir do conjunto mais extenso das frases do corpus.

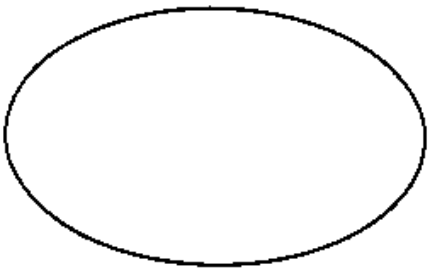
Em qualquer caso, a questão da representatividade deve ser bem colocada e profundamente reflectida. Um exame profundo dos textos ou dos dados linguísticos previamente armazenados em corpora informatizados é determinante do sucesso da selecção a realizar tendo em vista a constituição de um corpus, ou subcorpus, para um estudo específico.

Embora em Linguística Histórica não se possa recorrer à introspecção ou à eliciação para colmatar as insuficiencias dos textos antigos, é animador poder considerar que o conhecimento de diferentes gramáticas, em que se delimitam, por um lado, as características reguladas por princípios gerais, por outro, as particularidades associadas aos valores dos parâmetros universais, deverá favorecer a formulação de hipóteses e a análise explicativa dos dados atestados no corpus.

análise + explicação

fundamentadas teoricamente

constituem o



Referências

McEnery, Tony & Andrew Wilson (1996)
Corpus Linguistics, Edinburgh University Press, p. 59 (m/ trad. e realces coloridos).
Xavier, M.F.; A. Castro; A. Gonçalves (2001)
"A mais Antiga Terminologia Notarial Portuguesa" in *Actas do Congresso sobre a Língua Portuguesa no Brasil*, Universdidade de Évora (no prelo).
Xavier, Vicente e Crispim *orgs.* (1999)
Dicionário de Verbos Portugueses do Século 13, Lisboa, Centro de Linguística da Universidade Nova de Lisboa, Linha de Investigação 1.
Cf. Xavier, Vicente e Crispim *orgs.* (2002)
Dicionário de Verbos Portugueses Medievais dos Séculos 12 e 13/14, Centro de Linguística da Universidade Nova de Lisboa, Linha de Investigação 1.