Os Corpora Linguísticos: uma nova forma de "fazer Lexicografia"?

Ana Rebello de Andrade
ILTEC / FLUL / FCT / ISEC[1]

Introdução genérica

A linguística de corpora é um fenómeno relativamente recente na história da linguística. A constituição de um corpus electrónico era, há alguns decénios, uma árdua tarefa executada em centros de cálculo (cf. Habert, Nazarenko, Salem: 1997:143) por máquinas com capacidades de armazenamento e de cálculo limitadas.

Também a "revolução gerativa[2]" valorizando a noção de *competence* em detrimento da *performance* promoveu como metodologia, para a obtenção de dados linguísticos, a introspecção do próprio falante-linguista.

Com o advento da nova era de micro-informática e consequente aumento de capacidade e velocidade no armazenamento/processamento dos dados e a criação de redes informáticas, a situação alterou-se radicalmente possibilitando, com relativa facilidade, a emergência dos corpora linguísticos que, no entender de Sinclair, são "colecções" significativas de dados seleccionados e organizados segundo critérios linguísticos explícitos para cumprir determinadas funções. (cf. 1996:4)

Também no âmbito da própria gramática gerativa a obtenção de dados é, hoje em dia, vista de uma forma menos angulosa ou estreita, sendo os corpora usados para ilustrar os exemplos que decorrem das posturas teóricas assumidas por esta vertente da Linguística moderna.

Durante uma visita a Lisboa, a convite do CLUL (projecto DCP) e da FLUL (Departamento de Língua e Cultura Portuguesa), e numa das duas conferências proferidas[3] J. Sinclair discutiu a noção de *performance* e *competence* concebidas por Chomsky, equacionando-as em termos de **competence-potencial / performance--actual**. Como refere Neto (1995: 27), segundo Sinclair, "a existência potencial na língua de determinada estrutura não tem interesse, pela simples razão de que não

---

[1] A autora é doutoranda na *Faculdade de Letras da Universidade de Lisboa* (FLUL), tem como instituição de acolhimento o *Instituto de Linguística Teórica e Computacional* (ILTEC), é bolseira de doutoramento da *Fundação para a Ciência e Tecnologia* (FCT), com a ref.ª SFRH / BD/ 5442 /2001, e é Prof.ª Adjunta no *Instituto Superior de Educação e Ciências* (ISEC).

[2] Utilizo o termo **gramática gerativa,** consagrado no português do Brasil e não **gramática generativa,** variante utilizada no português europeu por achar mais correcta do ponto de vista morfológico (construção da palavra) a primeira variante – sendo a segunda, na minha opinião, uma tradução apressada do inglês do termo "generative".

[3] 16 de Janeiro no CLUL e 17 de Janeiro na FLUL.

***Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2002, pp. 103-110.**

existe. (...) Ao contrário, a dimensão ***actual*** (**real**, em português) é mensurável e existe realmente. É uma dimensão com um valor individual e com um valor social. Assim um corpus linguístico reflectiria a dimensão **real** na língua de uma sociedade."

Nesta comunicação pretende-se, num primeiro momento, elaborar uma pequena história dos corpora nos estudos linguísticos (nascimento, desenvolvimento, tipologia), sublinhando as vantagens e desvantagens da utilização/constituição de um corpus para análise linguística, para, num segundo momento, reflectir, através de exemplos concretos, sobre a sua importância na realização de produtos lexicográficos gerais e de especialidade.

História dos corpora nos estudos linguísticos

2.1. Nascimento e desenvolvimento[4]

Pelas razões anteriormente explicitadas, a constituição de corpora linguísticos para obtenção de dados, não tinha, até aos anos 50, um grande impacto. No entanto, nos finais dos anos 50, dois acontecimentos deram lugar à expansão dos corpora:

– a recolha de Randolph Quirk para um grande corpus de língua inglesa, falada e escrita, que veio a ser conhecido como o **SEU** corpus (*Survey of English Usage*);

– o advento dos computadores, com a primeira máquina para leitura e classificação de performances linguísticas de Nelson Francis e Henri Kucera (Brown University).

Quando Randolph Quirk anunciou, em 1959, os seus planos de recolha de um grande corpus de língua inglesa (50% escrita e 50% oral, o que não foi totalmente realizado devido a problemas na transcrição e processamento dos dados do oral) foi logo seguido por Nelson Francis e Henri Kucera, que, entre 1961 e 1964, constituíram, nos E.U.A o *Brown Corpus*, de um milhão de palavras.

Também em França, e a partir dos anos 50, surgiu a ideia original do Francês Fundamental, corpus que contém dois tipos de corpora complementares:

– um corpus de frequência formado por transcrições do oral;

– um corpus de disponibilidade baseado em inquéritos temáticos feitos em colégios, liceus e mais tarde junto de adultos, através das técnicas de elicitação e que dá conta de vocabulário adequado a temas de interesse geral, vocabulário que não ocorre com grande frequência no quotidiano.

A ideia deste corpus inspirou outros projectos congéneres que alguns anos mais tarde vieram a ser concretizados, tais como o do *Alemão Fundamental* (A Pfeffer, 1961), o do *Espanhol Fundamental* (P. Rivenc e A Rojo Sastre, 1969) assim como o do *Português Fundamental* que teve o seu início em 1970, projectos que foram patrocinados pelo Conselho da Europa.

---

[4] Para mais detalhe sobre este assunto, veja-se Rebello de Andrade (1995: p.25 e segs).

Também por volta dos anos 70, Jan Svartik concretizou a tarefa de tornar legíveis por computador os textos orais do **SEU** corpus. O resultado obtido foi o *London-Lund Corpus* (**LLC**) com mais de 500.000 de palavras e que, ainda hoje, é uma fonte procurada para o estudo da língua inglesa falada.

A esta "primeira geração" de corpora sucede-se uma "segunda geração" representada pela *Birmingham Collection of English Text*, orientada por J. Sinclair, que data dos anos 80, e que contém, no seu corpus principal, vinte milhões de palavras.

O avanço tecnológico da informática – concretamente na área de reconhecimento de caracteres – permitiu o processamento mais rápido dos dados, e consequentemente um alargamento quantitativo dos mesmos. Donde resulta que, actualmente, os corpora de "terceira geração" tenham largos milhões de palavras tal como o *Bank of English* (200 milhões de palavras), o *CREA* (corpus espanhol com 200 milhões de palavras) comparável ao *Frantext* de 240 milhões de palavras e que serviu de base à elaboração do *Trésor de la Langue Française* (T.L.F.). Tal como reitera Sinclair (1991: 25): "Não precisamos de estimar o texto em si próprio já que vivemos numa época de explosão textual"[5]

O *Corpus de Referência do Português Contemporâneo* (C.R.P.C).) é um projecto que teve o seu início em 1990 e está a cargo do Centro de Linguística da Universidade de Lisboa (CLUL). Este corpus inclui, actualmente, 152, 6 milhões de palavras e está previsto o seu aumento. Inclui amostragens de língua escrita e de língua falada.

2.2. Tipos de corpora

Existem diferentes tipos de corpora linguísticos. Sinclair (1991:23) divide-os em dois subtipos:

– os "sample corpora" – que se caracterizam por serem corpora de referência, onde as produções linguísticas neles incluídas são amostragens da língua falada e escrita, residindo o seu valor na forma como se combinam percentualmente os vários registos da língua;

– os "monitor corpora" – que se distinguem dos primeiros por ser dada primazia à quantidade sobre o "equilíbrio" (forma como se combina a quantidade e a variedade de produções linguísticas).

Dentro destes dois grandes géneros de corpora situam-se os *corpora de língua geral* e de *especialidade*, os *corpora comparados* e os *corpora paralelos*.

Os corpora de língua geral pretendem recensear, tal como o nome indica, a língua corrente nas suas várias realizações (do formal ao informal, do escrito ao oral, do quotidiano ao especializado, do literário ao coloquial).

Os corpora de especialidade pretendem, por sua vez, reproduzir, quantitativamente e representativamente a linguagem utilizada num determinado domínio do conhecimento humano (científico, técnico, artístico).

---

[5] Tradução livre da minha autoria: "We do not need to cherish text; we live in a time of textual explosion."

Quanto aos corpora comparados e paralelos podemos afirmar que estes são corpora construídos a pensar em estudos linguísticos comparados (sincrónicos e diacrónicos, de uma ou mais línguas) e são, consoante os objectivos a atingir, monolingues, bilingues ou plurilingues.

Os corpora comparados são constituídos a partir de fontes (autores), registos e dimensões, comparáveis entre si, tanto em termos de dimensão como em termos de variação – sendo certo que as produções neles incluídas não têm o dever de ser "idênticas", no sentido, por exemplo, de serem textos originais e suas respectivas traduções. Este factor torna-se pertinente, no caso de os corpora comparados poderem ser elaborados em mais de uma língua.

Os corpora paralelos são exclusivamente constituídos em pelo menos duas línguas, sendo as produções neles incluídas obrigatoriamente textos, provindos de traduções de uma mesma fonte ou documento.

2.3. Vantagens e desvantagens na utilização/constituição de um corpus para análise de dados linguísticos

Sendo um corpus linguístico uma rica fonte de dados ele não deixa de ser uma opção a tomar tendo em conta as suas vantagens e desvantagens na obtenção de dados para um trabalho linguístico de carácter descritivo de uma determinada área da língua ou da língua em geral.

Utilizar um corpus linguístico significa (cf. Neto: 1995: 33 e segs):

– poder utilizar um corpus já constituído e que pelas suas características sirva os objectivos da análise pretendida;

– constituir um corpus específico, "recortando-o" de um corpus já existente, seleccionando dentro de um corpus maior os documentos desejados de forma a constituir um sub-corpus;

– constituir um corpus a partir do zero, procedendo a todo o trabalho de desenho, planificação e recolha dos elementos.

Decorrente destes três tipos de situações enunciadas surgem as seguintes vantagens e desvantagens.

As vantagens na utilização de um corpus como método para a obtenção de dados atestados da língua "in vivo" prendem-se essencialmente com o facto de podermos aceder a um conjunto de dados reais e ricos no sentido em que se o corpus for extenso e variado aparecerão, de forma clara, as unidades de comunicação mais utilizadas e as menos utilizadas, assim como os seus padrões semânticos, as associações que estabelecem entre si (combinatórias fixas ou privilegiadas), as variações de cada unidade ou pelo menos das mais frequentes, entre outros.

As desvantagens na opção pelo uso de um corpus, tal como sublinha Neto (1995: 34 e seg.), articulam-se, basicamente, sob duas diferentes ordens de ideias:

Por um lado, a própria natureza e constituição do corpus, que só será representativo se for equilibrado em termos de dimensão e variação dos textos e registos nele incluídos. Caso tal não aconteça alguns inconvenientes poderão fazer perigar o rigor e a seriedade do estudo efectuado. Neto (1995: 35) refere que: "Uma dimensão insuficiente pode comprometer a validade das conclusões que se desejam extrair a partir da análise de dados, especialmente se se utilizarem métodos quantitativos, o que sucede com muita frequência no caso de corpora computorizados." Enquanto que uma variação insuficiente poderá não dar conta dos vários domínios do discurso científico a analisar.

Por outro lado, desenhar, planificar e construir um corpus é um processo que requer meios humanos, informáticos e temporais suficientes para o fazer – o que nem sempre é possível –, tratando-se de um processo muito oneroso tanto a nível financeiro como a nível de tempo investido. Por esta razão poderá não ser uma escolha comportável.

3. Importância dos corpora na elaboração de produtos lexicográficos

A própria noção de corpora veio mudar a forma de fazer Linguística, tendo-se, por um lado, demarcado enquanto disciplina relativamente autóctone da Linguística e, por outro lado, podendo-se assumir como opção metodológica. Trabalhar com um corpus linguístico ou a partir dele significa analisar os dados sem "preconceitos" ou hipóteses teóricas muito sofisticadas *a priori*, para formular, *a posteriori*, os caminhos teóricos baseados numa leitura e interpretação atentas da "dimensão real" da língua, impressa de valor individual e social.

De facto, os corpora têm desempenhado uma função decisiva no desenvolvimento de áreas da Linguística como a semântica, a análise textual, a história da língua e a sintaxe, mas é sobretudo na lexicografia que maior impacto têm tido, tendo conduzido a uma redefinição do próprio conceito de lexicografia, bem como a uma nova concepção do dicionário e da forma de o elaborar.

Tradicionalmente, os produtos lexicográficos, nomeadamente os dicionários de língua geral ou de língua especializada, são elaborados a partir da selecção de uma nomenclatura baseada em trabalhos congéneres mais antigos, em recolhas pacientes dos seus autores acerca do interesse e sentido de determinadas unidades lexicais e na intuição do lexicógrafo (responsável) ou da equipe lexicográfica encarregada da elaboração de um determinado produto.

Também, a nível da própria definição do artigo dicionarístico, esta é muitas vezes baseada nos sentidos atribuídos anteriormente àquela palavra, podendo vir a revelar-se desadequada sincronicamente falando, ou demasiado restritiva. O sentido de uma palavra, na lexicografia tradicional, corresponde, na maior parte das vezes, a um sentido inerente, não raro, prototípico, dessa mesma palavra, esquecendo, os sentidos usados "realmente".

Os exemplos e abonações de cada entrada de dicionário são, por sua vez, baseados na mundividência e intuição linguística do próprio redactor, ou então, retirados de textos literários ou outros de forma, por vezes, ocasional, originando, muitas vezes, exemplos artificiais. As abonações que enchem estes dicionários conferem à definição da própria palavra, usos que, no fundo, não dão conta da própria língua "em uso", já que são abonatórios de usos e registos mais formais e sofisticados.

O desdobramento polissémico que a maior parte das unidades lexicais assume numa língua – e este aspecto é particularmente pertinente no caso da língua geral – é também tratado, do ponto de vista lexicográfico tradicional, de uma forma menos próxima da realidade linguística, já que os vários sentidos para um item são baseados em critérios como a existência ou não desses sentidos em trabalhos congéneres, a própria intuição do redactor do artigo sobre a unidade em questão e a verificação "pontual" em textos (jornais, textos literários etc.).

Basear um produto lexicográfico num corpus significa "virar ao contrário" a forma tradicional de desenvolver este tipo de produto. Significa que:

Poderá ser escolhida uma nomenclatura baseada, nomeadamente, em critérios de frequência, ocorrendo as unidades mais correntes na comunicação geral, ou as menos correntes, já que muitas vezes o nosso conhecimento introspectivo nem sempre corresponde à realidade do uso da língua. A selecção de uma nomenclatura fundamentada em corpora tem a vantagem de ser mais realista, no sentido em que um corpus retrata uma língua "viva", nos seus usos (quantitativos e qualitativos). Refere Sinclair (1987: XVIII): "Que palavras devemos incluir num dicionário? (...) As palavras comuns são muito importantes. Mas o que é ainda mais importante e resulta da nossa investigação é o facto das palavras comuns exibirem diferentes tipos de uso e significados subtis."[6]

Poderá ser elaborada a definição de cada unidade lexical, consoante o uso e sentido real que lhe é atribuído, com base nas performances, representativas, de falantes de uma língua num determinado estado, minorando por isso definições e sentidos com um carácter mais idiossincrático. Nas linguagens de especialidade este factor é de grande importância já que as definições conceptuais, em termos de normalização através de um dicionário de especialidade, devem assumir um carácter relativamente "rígido" e consensual. Um corpus de especialidade ao fornecer-nos os contextos definitórios retirados de fontes de especialidade pode, efectivamente, ser uma contribuição muito importante.

Os exemplos poderão ser retirados do corpus, conferindo ao artigo um carácter mais "natural" e menos "artificial" no sentido em que não são exemplos "forjados", mas sim produtos de performances em contextos "naturais" de comunicação.

---

[6] Tradução livre da minha autoria: "Which words, then, should we put in a dictionary? (...)The common words are very important. What is more, our research shows that the common words have many different patterns of use and subtle range of meanings."

Os vários sentidos que uma palavra pode assumir, consoante o contexto comunicativo, são muito visíveis nos corpora, podendo assim estes representar um contributo importante neste aspecto.

A ordenação das acepções poderá ter um carácter mais sistemático, baseando-se na frequência de uso ou na cronologia.

Também a nível do tipo de associações semânticas e contextuais que as palavras estabelecem entre si e que vão desde a co-ocorrência privilegiada até à combinatória fixa, um corpus pode-nos ser de grande utilidade. Neto (cf. 95:171 e segs.) refere a utilidade do corpus ASTRO (corpus especializado de Astronomia) para a determinação da co-ocorrência de determinadas unidades em combinação com outras menos esperadas – tal como por exemplo, galáxias elípticas anãs.

As várias "funcionalidades" que uma unidade lexical exibe são também observáveis num corpus, podendo alterar a intuição linguística do lexicógrafo, *a priori*, acerca das prioridades em termos de categoria gramatical de uma determinada unidade. Um bom exemplo disto é a palavra "bom" que intuitivamente classificamos como adjectivo que é, tendo sido observada, numa primeira análise, no Corpus REDIP[7], uma alta frequência (80 ocorrências em 250) enquanto bordão linguístico.

Esta forma de "fazer lexicografia" gerou nas últimas décadas alguns produtos lexicográficos interessantes, nomeadamente:

English Language Dictionary da Editora Collins, dirigido por John Sinclair e elaborado pela equipa da Universidade de Birmingham (dicionário conhecido por Cobuild);

O Trésor de la Langue Française;

Também o *Dicionário de Língua Portuguesa Contemporânea* publicado pela editorial Verbo e elaborado pela equipa da Academia das Ciências, baseia as suas fontes no maior corpus português o CRPC, projecto desenvolvido pelo CLUL.

Tal como refere Malaca Casteleiro, na introdução (2001: XIV): "A maior parte destes recursos documentais foi facultada à Academia através do Corpus de Referência do Português Contemporâneo "

Notas conclusivas

Nesta comunicação pretendeu-se dar a conhecer uma breve história dos corpora procurando enfatizar a sua importância na forma como a linguística de corpora tem contribuído para o desenvolvimento de áreas da Linguística – nomeadamente a lexicografia.

[7] Este é um projecto em curso no ILTEC, em parceria com o CLUL e a Universidade Aberta e que visa a criação de um corpus de registos discursivos dos meios de comunicação e realização de descrições lexicais, gramaticais e textuais a partir desse corpus.

Referências bibliográficas:

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

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

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

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

English Language Dictionary 1987. London / Glasgow: Collins Cobuild.

Dicionário de Língua Portuguesa Contemporânea. 2001.Lisboa: Verbo, Academia das Ciências / Fundação Calouste Gulbenkian.