O USO DE *CORPORA* NA ANÁLISE LINGÜÍSTICA[1]

Guilherme Fromm[2]

> ***Corpus***: substantivo masculino; **1.** Coletânea ou conjunto de documentos sobre determinado tema. Ex.: ele estuda o c. juris canonici (coletânea de direito canônico); **2.** Derivação: por analogia. Repertório ou conjunto da obra científica, técnica e/ou artística de uma pessoa ou a ela atribuída. Ex.: o c. da poética camoniana; **3.** Rubrica: fisiologia. Estrutura com características ou funções especiais no corpo de um homem ou de um animal; **4.** Rubrica: lingüística. Conjunto de enunciados numa determinada língua, ger. colhidos de atos reais da fala, que servem como material para análise lingüística; **5.** Rubrica: lingüística, semiologia. conjunto de enunciados (que são indefinidamente possíveis, i.é., inesgotáveis), constituído por amostras significativas da gramática de determinada língua. Ex.: o c. sintagmático da língua portuguesa.
>
> Dicionário eletrônico Houaiss da língua portuguesa

O que é um *corpus*?

Percebemos, pelo exemplo acima, que *corpus* não apresenta somente um conceito. Com exceção das definições dois e três do verbete apresentado, todas as demais definições abarcam a nossa proposta que é definir, como montar e trabalhar com *corpora*. De um modo geral, *corpus*, na área da Lingüística, indica uma coleção de textos reunidos, de áreas variadas ou não, com um propósito específico de análise. Ele difere-se, portanto, de uma coletânea (coleção de trechos de obras) ou de uma antologia (uma coleção de textos de autores consagrados), que reúnem obras ou parte de obras dispersas com um intuito didático ou simplesmente comercial.

O uso de um *corpus* para validar resultados de uma pesquisa é um expediente utilizado há séculos, mas a ciência da Lingüística do *Corpus* (*Corpus Linguistic*) é relativamente nova. As conceituações na área ainda são bastante recentes e nem sempre aplicáveis para qualquer tipo de trabalho. Para nos atermos a um dentre vários conceitos

---

[1] Publicado originalmente na Revista Factus nº 1 (2003) – ISSN 1679-1851. FROMM, Guilherme . O USO DE CORPORA NA ANÁLISE LINGÜÍSTICA. Revista Factus, São Paulo, v. 1, n. 1, p. 69-76, 2003.

[2] Mestre em Lingüística e doutorando em Língua Inglesa pela FFLCH/USP; professor da Uniban.

possíveis do que é um corpus, optamos por recorrer à definição proposta por BIDERMANN (2001, p. 79):

> ... *corpus* constitui um conjunto homogêneo de amostras da língua de qualquer tipo (orais, escritos, literários, coloquiais, etc.). Tais amostras foram escolhidas como modelo de um estado ou nível de língua predeterminado. A análise dos dados lingüísticos de um *corpus* deve permitir ampliar o conhecimento das estruturas lingüísticas da língua que eles representam.

BIDERMANN (idem) fornece, ainda, uma segunda concepção de corpus:

> Pode-se definir um *corpus* lingüístico informatizado assim: - é uma coletânea de textos selecionados segundo critérios lingüísticos, codificados de modo padronizado e homogêneo. Essa coletânea pode ser tratada mediante processos informáticos.

BAKER (1995, p. 229) já trabalha com a composição interna e nos apresenta alguns critérios na seleção de um corpus:

> Corpora are generally designed on the basis of a number of selection criteria, the most important of which are:
> (i) general language vs. restricted domain
> (ii) written vs. spoken language
> (iii) synchronic vs. diachronic
> (iv) typicality in terms of range of sources (writers/speakers) and genres (e.g. newspaper editorials, radio interviews, fiction, journal articles, court hearings)
> (v) geographical limits, e.g. British vs. American English
> (vi) monolingual vs. bilingual or multilingual.

Essas são algumas concepções, dentre várias. Em virtude de ser uma ciência incipiente (década de 80 em diante, com o advento do computador como um instrumento acessível a muitos e a possibilidade de trabalhar com bancos de dados), a Lingüística do *Corpus* ainda não possui definições rígidas de seu objeto de estudo e nem de suas metodologias.

O que fazer com um *corpus*?

Antes de esquematizarmos um projeto para um corpus, devemos ter em mente quais aplicações práticas queremos retirar desse *corpus*. Existe uma variada gama de análises lingüísticas que podemos fazer a partir dele. Vejamos alguns exemplos:

- A freqüência das palavras mais comuns da língua;
- a freqüência das classes gramaticais;

- comprovação de colocações[3] na língua;
- reconhecimento e detalhamento de lexias compostas e complexas[4];
- regência dos verbos preposicionados;
- composição mais provável das estruturas frasais cristalizadas, tais como os provérbios e expressões idiomáticas;
- seleção de uma nomenclatura para uma obra terminológica;
- criação de dicionários gerais multilingües;
- verificação de modalidades de tradução em corpus bilíngüe ou multilingüe;
- base de dados para tradutores;
- ensino de língua estrangeira.

Definindo o objetivo

A primeira etapa, antes da coleta do material, é a indagação acerca dos objetivos que esperamos alcançar: que tipo de pesquisa pretendemos aplicar nesse *corpus*? Para quem se destina esse *corpus*? Quais são as fontes a serem trabalhadas? Que tamanho pretendemos para esse *corpus*? Em qual meio (escrito ou eletrônico) ele deverá ser publicado?

Todas essas perguntas, formuladas previamente, nos ajudam a estruturar o corpus e, acima de tudo, nos ajudam a economizar tempo (já que provavelmente lidaremos com grandes quantidades de informação).

A criação de um *corpus* de análise - exemplificação

As idéias apresentadas até agora podem nos dar uma pista de como trabalhar com a *Lingüística do Corpus*, mas acreditamos que elas seriam vagas se não

---

[3] A colocação indica uma combinação provável mais aceita pelos falantes nativos da língua. Ela é arbitrária e não segue padrões pré-estabelecidos pela semântica ou sintaxe. Ex.: *tomar um ônibus, pegar um ônibus, agarrar um ônibus*. Embora *agarrar* possa ser considerado, em determinado nível conceptual, como sinônimo de *tomar* ou *pegar*, provavelmente só encontraríamos os dois primeiros exemplos em uma análise de um *corpus*. O terceiro exemplo, embora possamos entendê-lo, nos soa estranho, não é uma colocação usual da língua.

[4] As lexias compostas seriam geradas por duas palavras (justapostas ou hifenizadas), gerando uma terceira palavra e um terceiro sentido, porém ainda guardando uma relação de significação com os dois sentidos originais. Ex.: *guarda-chuva*. As lexias complexas trabalham no nível frasal, onde podemos ou não recuperar o sentido original de cada lexia simples, mas o sentido final é independente delas. Ex.: *certificado de depósito bancário* (CDB).

mostrássemos um exemplo para ilustrá-las. Para tanto, usaremos um exemplo tirado de nossa dissertação de mestrado (FROMM, 2002): a construção de um *corpus* de informática. O objetivo desse foi fornecer palavras e suas respectivas exemplificações, na área de informática, baseadas em um critério de freqüência, para a construção de um glossário terminológico para tradutores.

Conteúdo

A delimitação da área de pesquisa desse *corpus* centrou-se em publicações gerais (revista e jornais via Internet) sobre a área de informática. Excluímos, portanto, publicações específicas (como revistas voltadas exclusivamente para programadores, com linguagens de programação e termos estritamente técnicos), manuais e *press-release* de companhias (que representam o uso de termos específicos desta ou daquela companhia) e outros por acreditarmos que não ofereceriam uma abrangência relevante para a formação de um glossário geral.

Decidimos, num primeiro momento, não especificar as subáreas apresentadas nas publicações (hardware, software, rede, dúvidas de leitores, dicas técnicas, etc.) por acreditarmos que os termos eram usados por todas elas quase que indistintamente.

Há, obviamente, diferenças quando tratamos de textos escritos em português com empréstimos do inglês e traduções de artigos do inglês em que se optou por não traduzir ou "decalcar" tal termo. Tais diferenças, quando não devidamente observadas, poderiam provocar conclusões errôneas há alguns anos (em que grande parte do material publicado era traduzida); percebemos, porém, dentro do *corpus* proposto, que as traduções não chegavam a 5% do total. Ainda assim, decidimos separar o material escrito em português das traduções.

Origem dos textos

O projeto inicial da pesquisa previa a coleta de material a partir de cinco publicações: cadernos de informática dos jornais **O Estado de São Paulo** e **Folha de São Paulo**, além das revistas **INFO Exame** (Editora Abril**), PC Master** (Editora Europa) e **Internet.br** (Ediouro), totalizando cinco diferentes fontes de análise no período de um ano (iniciando-se em janeiro/2001). Os cadernos de informática dos dois jornais e o conteúdo da revista INFO eram disponibilizados na Internet, portanto de fácil coleta via download. As outras duas revistas não eram disponibilizadas na Internet,

razão pela qual seriam escaneadas e compiladas em formato digital. Após alguns meses de coleta do material, chegamos à conclusão de que seria muito difícil escanear tantas revistas, dada a deficiência dos programas de OCR (reconhecimento ótico de caracteres).

Decidimos, então, que o material deveria vir totalmente da Internet. Portanto, a configuração final, determinou a inclusão dos cadernos de informática dos jornais **O Estado de São Paulo** e **Folha de São Paulo** (semanais) e a revista **INFO Exame** (mensal)[5]. Todo o material foi coletado entre janeiro e dezembro de 2001.

Identificação do Corpus

Tendo em vista que uma separação por fonte não era o objetivo da análise (já que não procurávamos linguagens específicas desta ou daquela editora e sim termos gerais da área), o critério que melhor se ajustou à organização do material foi a ordem cronológica. As fontes foram relacionadas da seguinte maneira (acompanhadas na seqüência das datas de publicação e separadas por mês):

FSP: Folha de São Paulo
OESP: O Estado de São Paulo
INFO: INFO Exame

Cada arquivo continha o texto integral de uma edição (com exceção das propagandas, gráficos, charges e tabelas de preços[6]). Procedemos ainda, a uma separação entre os textos originalmente escritos em português e as traduções. Resultou desse processamento três sub-*corpus*:

1. *Corpus* geral: todo o material recolhido em todas as publicações com divisão cronológica; salvo em formato .doc (Word). Exemplo: FSP 04.04.2001.doc, contido na pasta *Corpus Geral Word/Abril*.

[5] Essas publicações disponibilizavam o material na íntegra. Muitas outras foram cogitadas como fontes, mas, por fornecerem somente trechos e/ou algumas reportagens da versão impressa, logo foram abandonadas.

[6] Tal exclusão se deu mais por uma necessidade prática do que por uma escolha metodológica. Tais seções poderiam enriquecer ainda mais o corpus, porém, devido à diagramação ou o formato gráfico, elas não podem ser lidas e/ou convertidas para o formato final de leitura do conjunto.

2. *Corpus* português: todo o material escrito em português, com divisão cronológica, no formato .txt (texto). Exemplo: OESP 01.01.2001.txt, contido na pasta *Corpus Português/Janeiro*.
3. *Corpus tradução*: todo o material traduzido (FSP, OESP e INFO), com divisão cronológica, no formato .txt (texto). Exemplo: INFO.txt, contido na pasta *Corpus Tradução/Maio*.

Três motivos fizeram com que esse tipo de divisão fosse efetivada:

1. o programa que faz a contagem e separação das palavras de acordo com a freqüência, o WordSmith Tools, só aceita documentos no padrão .txt para análise;
2. a separação entre documentos originalmente escritos em português e traduções facilitaria um futuro etiquetamento[7] desses textos e a aglutinação dessas amostras dentro de um corpus geral da língua;
3. já que no formato .txt perde-se toda a formatação, resolvemos salvar os textos integrais com a formatação original; dessa forma, seria facilitado o entendimento dos textos e salvaguardado o acesso de algum pesquisador que necessitasse separá-los pelo título.

Dada a inviabilidade prática da impressão dessas amostras devido ao tamanho, eles foram salvos em CD-ROM, em pastas e sub pastas conforme explicado acima

Extensão/Dimensão

No total, organizamos e analisamos 52 edições dos cadernos de informática de ambos os jornais, além de 12 edições da revista INFO. Como resultado da contagem final das palavras, apuramos os seguintes números:

Tokens[8]: 1.392.706

Types: 48.482

[7] Pode-se *etiquetar* um corpus, ou seja, classificar morfologicamente e/ou sintaticamente palavra por palavra do mesmo. Já existem tentativas de criar programas que façam esse tipo de trabalho automaticamente, mas de um modo geral ainda é um serviço braçal.

[8] "Na língua inglesa os estatísticos do léxico costumam opor o *token* (ocorrência no texto) ao *type* (lexema referido pela ocorrência formal)." (BIDERMAN, 2001, p.167)

SARDINHA (inédito, 1999), propõe que se classifique os *corpus* segundo o número de palavras contidas:

Tamanho em Palavras	**Classificação**
menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio Grande
10 milhões ou mais	Grande

Seguindo esse critério, o autor chega à segmentação em cinco níveis de tamanho. Com base nesse critério de medição, pode-se afirmar que o conjunto de material selecionado para esta pesquisa constituiu um *corpus* de tamanho médio-grande. Isso permite dizer que era um *corpus* representativo no universo das revistas e jornais na área de informática.

Concluindo

A construção de um *corpus* ou *corpora*, gerais ou específicos, requer um grande planejamento prévio por parte do pesquisador. A falta desse poderá invalidar os dados obtidos na futura pesquisa. Cabe ao(s) pesquisador(es), devidamente aparado nas pesquisas metodológicas mais modernas dentro da área, desenvolver esse planejamento.

O desenvolvimento do *corpus*, dependendo do tamanho (quanto maior, mais representativo ele será), requer a participação de vários pesquisadores e auxiliares e pode demorar anos para ser terminado. Existem ainda muitos *corpora*, financiados por grandes instituições governamentais ou particulares, que não tem um fim planejado: eles são continuamente alimentados com novos dados para servirem como base de pesquisas diversas, tendo elas sempre um caráter de atualidade em relação à língua vigente.

Tendo em vista esses detalhes, acreditamos que há a real possibilidade de construção de teorias lingüísticas baseadas em fatos, passíveis de serem re-analisados. Essas teorias não se baseariam em “soluções” (ou mais especificamente, exemplificações) criadas por autores, mas sim em colocações autênticas da língua em estudo, conferindo-lhes um caráter científico.

Bibliografia

BAKER, M. Corpus in Translation Studies: an overview and some suggestions for future research. In: **Target 7:2**. Amsterdam: John Benjamins, 1995.

BIDERMANN, M.T.C. **Teoria Lingüística**. 2. ed. São Paulo: Martins Fontes, 2001.

FROMM, G. **Proposta para um modelo de glossário de informática para tradutores**. Dissertação de Mestrado. São Paulo: FFLCH/USP, 2002.

HOUAISS, A. **Dicionário Eletrônico Houaiss da Língua Portuguesa**. São Paulo: Objetiva, 2001.

SARDINHA, T. B. **O que é um corpus representativo?** Inédito, 1999.