JUDITH C. HOFFNAGEL
Universidade Federal de Pernambuco (UFPE)

LÍNGUA FALADA E LÍNGUA ESCRITA: A FORMAÇÃO DE CORPUS

Abstract

This paper discusses the problem of constructing a ***corpus*** *adequate for a variety of linguistic investigations of speech and writing. Among the questions addressed are:* ***corpus*** *compostion (which texts should be included and the problem of representativeness);* ***corpus*** *size (size of each text and quantity of each text variety).*

Palavras-chave: *1) corpus; 2) fala e escrita; 3) gêneros textuais*

Introdução

Nesta retrospectiva do Projeto Integrado, ***Fala e Escrita: Características e Usos***, apresentarei um esboço das pesquisas feitas por mim nestes últimos seis anos, seguido de uma breve apresentação do projeto de pesquisa proposto para os próximos dois anos.

Os projetos e seus resultados

O primeiros quatro anos foram dedicados ao estudo da *modalização* na fala e na escrita. Os estudos iniciais buscavam determinar: 1) onde, na fala ou na escrita, encontravam-se os modalizadores epistêmicos; 2) quais os tipos de modalizadores, entre os vários recursos lingüísticos disponíveis, mais usados na fala e na escrita; e 3) quais os gêneros textuais na fala e na escrita que são mais ou são menos modalizados, e 4) quais as funções desempenhadas pela modalização. Num aspecto mais quantitativo constatamos: 1) que os modalizadores epistêmicos são duas vezes mais freqüentes na fala do que na escrita; 2) que na fala a forma modal mais freqüente são os verbos de atitude proposicional (*eu acho que, eu sei que, eu acredito que*, etc.) enquanto na escrita são os advérbios e sintagmas preposicionadas em função adverbial; e 3) que tanto na fala como na escrita há gêneros textuais que são mais modalizados, outros menos e alguns gêneros que não apresentam nenhuma modalização.

Com respeito às funções da modalização, as análises mostraram que não há diferenças nas funções desempenhadas pela modalização nos vários gêneros textuais. É mais uma questão relacionada à natureza da interação: diferentes situações de interação requerem um maior ou menor uso da modalização. Por exemplo, a análise detalhada do modalizador epistêmico mais usado na fala, *eu acho que*, demonstrou que contextos discursivos nos quais as relações interpessoais estão em primeiro plano requerem a modalização para manter a interação, ou seja, muitas vezes é uma questão de polidez. Também, nos textos em que o falante não tem nenhum ou tem pouco conhecimento do assunto em pauta, há uma tendência maior a modalizar. Outra investigação mostrou o papel da modalização na expressão irônica, especialmente nos textos de opinião política em jornais e revistas e o efeito humorístico que o uso da modalização freqüentemente provoca. A análise do gênero textual *bula de remédio* revelou um aspecto nocivo do uso da modalização. O uso adequado ou inadequado dos modalizadores modifica a força dos enunciados especialmente em sua função de advertência sobre os possíveis efeitos colaterais dos remédios.

Os resultados dos estudos sobre modalização sugerem que não são as poucas diferenças encontradas no uso das modalidades falada e escrita que são importantes ou interessantes, mas, as diferenças dos modalizdores na relação vista entre os gêneros textuais, falados e escritos. A modalização epistêmica e deôntica estão presentes na fala e na escrita, mas seu maior ou menor uso depende do gênero textual mais do que a modalidade da língua. O fato de constatarmos uma diferença no uso da modalização epistêmica e deôntica na fala e na escrita, sendo a epistêmica mais evidente na fala e a deôntica na escrita, provavelmente é devido, em grande parte, ao *corpus* utilizado nestas investigações. Por um lado, era relativamente restrito no tamanho e, por outro, limitado quanto aos diferentes gêneros textuais representados.

O projeto ora em andamento, entitulado *A Emergência de identidades sociais na atividade discursiva falada e escrita*, investiga os recursos lingüísticos-discursivos utilizados para a construção, manutençaõ e projeção de identidades sociais direta

ou indiretamente, tanto na exibição da própria identidade do falante/escritor como na sua atribuição de identidade a outros.

Há poucos recursos lingüísticos específicos que sempre são usados para indicar identidades sociais. Também, não há diferenças aparentes nas estratégias de identificação usadas nos textos falados e escritos. Há, no entanto, certos gêneros textuais, tanto na fala como na escrita, em que as identidades sociais são mais evidentes. Nos textos interativos, conversações, cartas pessoais, entrevistas, inquéritos judiciais e textos persuasivos (propaganda, artigos de opinião), por exemplo, há quase sempre uma expressão ou projeção de identidades sociais. Por outro lado, em textos de instruções, avisos, documentos governamentais, textos didáticos e científicos há pouca ou nenhuma expressão de identidade social.

Nas investigações sobre identidade social, como nas de modalização, parece ser mais pertinente o gênero textual e não a modalidade falada ou escrita que determina as variações.

O novo projeto

A pesquisa que está sendo proposta, *Princípios e critérios para a construção de um 'corpus de uso' para a análise lingüística da fala e da escrita,* tem suas raízes na constatação de que o *corpus* montado durante os seis anos de atuação do Projeto Integrado: Fala e Escrita se mostra cada vez mais insatisfatório. Uma das metas do Projeto Integrado desde seu início em 1995 foi a construção de um *corpus* de textos de uso real da língua falada e da língua escrita. Apesar das boas intenções dos pesquisadores envolvidos, não chegamos a formular critérios bem definidos para a construção e organização do *corpus*. Basicamente, duas decisões foram tomadas com referência à formação do *corpus*: (a) incluir uma variedade de gêneros textuais e (b) serem textos de uso real. O resultado é um *corpus* hoje com pouca variedade nos gêneros da fala e, embora tenha maior diversidade na escrita, há poucos textos na maioria dos gêneros representados no *corpus* e muitos textos em dois ou três gêneros.

Hoje, com o interesse cada vez mais acentuado nas investigações lingüísticas que enfocam o uso da língua na vida cotidiana e visando a aplicações dos resultados destas investigações em áreas como a educação, faz-se necessário repensar o que pode servir como dados para investigações. O que significa, por exemplo, dizer que um estudo foi baseado num *corpus* de dados reais? O problema é tanto teórico quanto metodológico. O problema central, então, para o projeto, é analisar princípios, critérios e condições para a sistematização de um *'corpus de uso'* dentro da perspectiva dos gêneros textuais.

Os estudos de uso tipicamente têm dois principais objetivos: 1) avaliar a extensão de uso do padrão sendo investigado e 2) analisar os fatores contextuais que influenciam sua variabilidade. O ponto crucial deste tipo de investigação, porém, é que para encontrar padrões de uso e analisar os fatores contextuais, precisamos de uma grande quantidade de materiais lingüísticos coletados de muitos e diferentes falantes/escritores para ter certeza de que não estamos baseando nossas conclusões nas idiossincrasias de poucos falantes/escritores. E esta necessidade de grande quantidade de linguagem traz dificuldades teóricas e metodológicas adicionais, no sentido de que as decisões metodológicas tomadas quando se constrói um *corpus* são baseadas em noções teóricas sobre a natureza do estudo lingüístico, a concepção de língua, a natureza dos dados lingüísticos, a natureza do comportamento lingüístico e o conceito de representatividade, entre outras.

Assim, antes de construir um *corpus,* precisam ser estabelecidos princípios e critérios que respondam a perguntas tais como: Quais são os tipos de investigação lingüística que pretendemos realizar com base nos dados do *corpus?* Qual é o tamanho ideal (se é que há um tamanho ideal) do *corpus* que permite ao investigador generalizar sobre a existência e variabilidade dos padrões de uso? Quais são os textos falados e escritos que devem ser incluídos? Existem alguns textos mais adequados para este tipo de investigação lingüística? Devem ser textos completos ou podem ser parciais? Devem estes textos representar todos os gêneros textuais que existem? (O que implica duas outras perguntas possíveis: o que é um gênero textual? e, sabemos quais são os gêneros textuais que existem?). O *corpus* deve servir a estudos sincrônicos e diacrônicos? Como organizar os dados coletados? Qual o tipo de organização/ sistematização necessária para as análises pretendidas?

Com estas questões em mente é que colocamos três objetivos básicos para a investigação em pauta:

(a) construir os critérios para a montagem de um corpus lingüístico controlado,
(b) definir os critérios para a elaboração de um programa computacional adequado ao tratamento dos dados de acordo com as possíveis variáveis de uma ampla gama de investigações lingüísticas atuais e futuras e
(c) possibilitar uma coleta sistemática e controlada de dados representativos dos mais diversos gêneros textuais.

O que é um corpus?

"Tradicionalmente", diz Leech (1997:1) "lingüistas usaram o termo *corpus* para designar um corpo de dados lingüísticos autênticos (que ocorrem naturalmente) que podem ser usados como uma base para a pesquisa lingüística". Biber et al. (1998:12) define o *corpus* como uma grande e metódica coleção de textos naturais.

Discutindo alguns dos problemas no desenho de um *corpus*, Biber et al. (1998:246) nos lembra que:

> *Um* corpus *não é simplesmente uma coleção de textos. Antes um* corpus *procura represen-*

tar uma língua ou alguma parte de uma língua. O desenho apropriado para um corpus*, portanto, depende do que se quer que ele represente. A representatividade do* corpus*, por sua vez, determina os tipos de problemas que podem ser pesquisados e a possibilidade de generalizar os resultados da pesquisa.*

A Representatividade

É importante frisar que a representação de uma língua – ou até parte de uma língua – é uma tarefa problemática. Não sabemos a extensão da variação em línguas ou todas as variáveis que precisam ser cobertas para captar toda a variação em textos. Contudo, a atenção a certos problemas deverá assegurar que um *corpus* será tão representativo quanto possível, dado nosso conhecimento corrente da língua.

Entre os gêneros textuais que raramente são incluídos em *corpora* mas que fazem parte da prática discursiva da maioria das pessoas, estão bilhetes, avisos, listas, instruções, formulários, receitas médicas, resultados de exames laboratoriais, boletins escolares, e propaganda. Sinclair (1991:18) nota que com a disponibilidade de jornais em forma eletrônica hoje em dia, se torna fácil incluir este material num *corpus*, mas o autor lembra que a língua de jornais é apenas uma variedade, ou grupo de variedades relacionadas, e não uma amostra representativa da língua.

Se estamos querendo construir um *corpus* representativo da língua portuguesa, por exemplo, como podemos decidir o que incluir? Biber et al. (1998:247) discutem a idéia da amostra proporcional. Isto é o *corpus* seria construído na base do uso da língua no cotidiano das pessoas. No entanto, como os mesmos autores apontam, um *corpus* proporcional deste tipo seria pouco útil para os estudos de variação, porque a maior parte do *corpus* seria relativamente homogênea. Isto é, num *corpus* proporcional ao uso da língua, a maioria dos textos seriam conversações e seriam muito semelhantes nas suas características lingüísticas.

A Diversidade de Textos

Uma preocupação particularmente importante para o desenho do *corpus* é a diversidade. Vários estudos já mostraram que há importantes diferenças no uso de elementos lexicais, gramaticais e discursivos através das diferentes variedades lingüísticas. Como Biber et al. (1996:248) apontam, não há, na verdade, algo que pudesse se chamar "linguagem geral"; cada registro tem seus próprios padrões de uso. Assim qualquer *corpus* que é usado para estudos de variação ou que procura representar uma língua precisa se preocupar com a diversidade de textos que inclui (ou exclui).

A própria noção de diversidade, no entanto, precisa ser esclarecida. Por exemplo, pode-se entender diversidade em termos das diferentes modalidades de língua (oral, escrita, etc.). Outro tipo de diversidade se refere à variedade de gêneros textuais (ficção, conversações, entrevistas, textos científicos, cartas pessoais, notícias televisivas, etc.). A diversidade pode ser entendida também em termos da variação dialetal-socioletal que existe para uma língua (variações regionais, variações por classes sociais, etc.)

Se é mais ou menos pacífico que um *corpus* deve incluir uma diversidade de textos, não é pacífico como deve ser organizada, classificada ou categorizada esta diversidade. Isto por várias razões, mas uma das mais importantes é a falta de uma clara distinção entre vários rótulos (variedade, registro, gênero textual ou discursivo, tipo textual, entre outros) usados para classificar os textos. Assim antes de escolher os textos para o *corpus* deve-se ter clareza sobre como as categorias são definidas.

Um estudo recente de Marcuschi (p.18 no prelo), *Gêneros Textuais: o que são e como se constituem*, ajuda estabelecer algumas distinções úteis para a construção de um *corpus* representativo da diversidade da língua portuguesa. Ele distingue três noções centrais:

a <u>*Tipo textual*</u>: é um construto teórico que abrange, em geral, de cinco a dez categorias, designadas *narração, argumentação, exposição, descrição, injunção.* Trata-se de um agrupamento pela natureza lingüística do texto produzido. Mais do que textos concretos e completos, estas são designações para seqüências típicas. Os tipos textuais não têm uma existência real.

b <u>*Gênero textual:*</u> é uma forma textual concretamente realizada e encontrada como texto empírico. O gênero tem uma existência real que se expressa em designações diversas, constituindo em princípio listagens abertas tais como: *telefonema, sermão, carta pessoal, romance, bilhete, etc***.** São formas textuais estabilizadas, histórica e socialmente situadas. Sua definição não é lingüística, mas de natureza sócio-comunicativa.

c <u>*Domínio discursivo*</u> (por exemplo: *Discurso jurídico, Discurso jornalístico, Discurso religioso,* etc.) não forma uma classificação de textos mas indica *instâncias de formação discursiva*. Constituem práticas discursivas mais amplas dentro das quais podemos identificar um conjunto de gêneros textuais.

Adotando estas noções teóricas, então, o *corpus* deve incluir uma diversidade de gêneros textuais, uma vez que estes são os únicos empiricamente realizados. Em outras palavras, os gêneros textuais representam práticas sociais.

Se por um lado parece razoável, então, decidir que o *corpus* será composto de uma variedade de gêneros textuais, por outro, não parece fácil chegar a uma definição de quais gêneros serão incluídos. Vários critérios de escolha podiam ser considerados: (a) os gêneros *mais praticados*, (b) os gêneros *mais comuns nos diversos domínios discursivos*, (c) os mesmos gêneros que *compõem os grandes corpora que existe para outras línguas*, etc. Cada uma destas

escolhas teria seus pontos positivos e negativos. Para (a) e (b), por exemplo, um ponto negativo seria conseguir definir 'mais praticado' e 'mais comum'. Em termos de produção (falar ou escrever) ou uso (ouvir ou ler)? Sem dúvida, para a fala, a conversação é o gênero textual mais comum, tanto na produção quanto no uso, mas não é tão fácil achar um gênero com estas características na escrita. Para (c), um ponto positivo seria a possibilidade de fazer comparações através das línguas.

O Tamanho do corpus

Qual o tamanho ideal para um *corpus?* A questão de tamanho não é fácil de resolver teórica ou praticamente. Como Biber et al. (1998:248-49) nota, freqüentemente discussões do tamanho do *corpus* focam exclusivamente no número de palavras do *corpus*. Mas, o problema de tamanho também se refere ao número de textos de diferentes categorias ou gêneros, o número de amostras de cada texto, e o número de palavras em cada amostra.

Edwards (1993:267) nota que enquanto há pouco tempo um *corpus* de um milhão de palavras foi considerado grande, hoje há projetos que procuram construir *corpora* de centenas de milhões de palavras. Para a língua escrita isto é relativamente fácil, hoje em dia, dado a disponibilidade de textos já informatizados (jornais, livros, etc.) bem como a facilidade de informatizar textos escritos através do uso do *scanner*, e assim eliminar a necessidade de re-digitar os textos. A língua falada, porém, é mais problemática. O processo de coleta dos dados orais e sua transcrição é extremamente demorado, aumentando, portanto, o custo da construção de *corpora* orais.

Sinclair (1991:18) oferece o seguinte conselho sobre tamanho do *corpus*: "[...] um *corpus* deve ser tão grande quanto possível e deve continuar a crescer." Com isto em mente é que pretendemos, após de estabelecer os princípios e critérios mencionados acima, construir um *corpus* da língua falada e da língua escrita que contém uma variedade de gêneros textuais autênticos. Será tão grande quanto possível e será desenhado para incorporar mais gêneros textuais na medida que novas práticas sociais produzem gêneros textuais.

Referências Bibliográficas

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge, Cambridge University Press.

Biber, Douglas; Conrad, Susan; e Reppen, Randi. 1998. *Corpus Linguistics: investigating language structure and use*. Cambridge, Cambridge University Press.

Edwards, Jane A. 1993. Survey of Electronic Corpora and Related Resources for Language Researchers. *In* Jane A Edwards and Martin D. Lampert, eds. *Talking Data. Transcription and coding in discourse research*. Hillsdale, N.J., Lawrence Erlbaum, pp. 263-306

Leech, Geoffrey. 1997. Introducing corpus annotation. *In* Roger Garside, Geoffrey Leech e Anthony McEnery, eds. *Corpus Annotation. Linguístic information from computer text corpora.* London, Longman, pp. 1-18.

Marcuschi, Luís Antônio. (no prelo) *Gêneros Textuais: o que são e como se constituem.* Recife.

Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation.* Oxford, Oxford University Press.