4 Estratégia

A Linguística de Corpus é uma abordagem empírica do estudo da linguagem, que, como fazem as ciências naturais há muito tempo, busca resultados através da observação de grandes quantidades de dados (Stubbs 2007, p. 16). O próprio Stubbs afirmou anteriormente:

A maior questão intelectual nas ciências sociais é a relação entre micro e macro. Como é possível que o comportamento rotineiro diário, de momento a momento, possa criar e manter instituições sociais ao longo de períodos extensos de tempo? (...) A análise de texto e corpus fornece métodos para o estudo das bases empíricas da sociedade nessas relações micro-macro.

(Stubbs 1996, p. 21)

A pesquisa em Linguística de Corpus apresenta quatro características, descritas por Conrad (2002, p. 76-78): o uso de um corpus, que é uma coleção de textos reais, originalmente escritos ou transcritos de originais falados, armazenados de forma eletrônica; a utilização de técnicas de análise que envolvem o uso de computadores, para dar conta da grande quantidade de dados; a ênfase na análise de padrões empíricos do uso da língua, com base nas descobertas feitas a partir do corpus; e o uso simultâneo de técnicas quantitativas e qualitativas de análise, mesmo que com o predomínio de um ou de outro aspecto. Além disso, diversas características interativas podem ser analisadas simultaneamente através de uma abordagem baseada em corpus (id, p. 82).

A pesquisa baseada em corpus permite assim descobrir, verificar e analisar características do uso da língua a partir de ocorrências reais. Por isso, o corpus é uma fonte primordial de informação acerca da linguagem (Sardinha 2000, p. 352).

Estudos baseados em corpus fornecem informações complexas sobre fatores sociais e textuais que influenciam as escolhas linguísticas, e portanto podem contribuir muito para nossa compreensão do discurso.

(Conrad 2002, p. 75)

4.1 Linguística Sistêmica e Linguística de Corpus

O presente trabalho utiliza a base teórica da Linguística Sistêmico-Funcional (LSF), bem como pressupostos da Linguística de Corpus. Embora nem sempre concordem, as duas correntes apresentam pontos em comum suficientes para permitir que sejam utilizadas de forma complementar.

O principal aspecto em comum, entre a LSF e a LC, como apontam Thompson e Hunston (2007, p. 4), é a preocupação com o estudo da língua em sua ocorrência natural e com a linguagem como texto. Além disso, como assinala Oliveira (2009, p. 5), ambas as áreas atribuem ao contexto situacional um papel determinante tanto na produção quanto na análise textual. Isso inclui os pesquisadores tanto da LSF quanto da Linguística de Corpus entre aqueles que acreditam que o trabalho do linguista é fundamentalmente concentrado na investigação do uso da língua. Segundo Thompson e Hunston (2007, p. 4), a noção de contexto situacional é um desafio para os linguistas que trabalham com corpus, já que os textos são despidos de contexto quando inseridos em um corpus. Por vezes, mas nem sempre, é possível recuperar parte do contexto situacional e/ou cultural, no caso de corpora que incluem informações ou classificações complementares (acerca de assunto, época de produção dos textos, autores ou participantes); mas a Linguística de Corpus parece estar mais preocupada com a identificação de padrões do que com a descrição de usos particulares em situações específicas, portanto, isso não parece ser um problema teórico relevante para a área (cf. Oliveira 2009).

Já a preocupação com a frequência, ao contrário, tem sido mais explícita na pesquisa com Linguística de Corpus que na LSF. Muitas das descobertas centrais da Linguística de Corpus têm sido relacionadas com a frequência de itens lexicais; além disso, linhas de concordância costumam revelar padrões típicos da língua que não são vistos quando exemplos isolados da língua são examinados (Thompson e Hunston 2007, p. 5). Vale ressaltar que esses padrões existem objetivamente; a técnica de observação os torna visíveis, não os cria, e por isso é possível fazer previsões a partir da análise de um corpus que são aplicáveis a outros corpora (Stubbs 2007, p. 18). A LSF, por sua vez, por ver a linguagem como potencial de significado, tem se preocupado mais com a probabilidade que

com a frequência relativa em si, ou seja, a frequência de ocorrência em exemplos é considerada uma realização das probabilidades potenciais do sistema (Thompson e Hunston 2007, p. 5). A escolha entre, por exemplo, ‘adeus’ e ‘tchauzinho’ é influenciada por fatores contextuais como a relação entre o falante e o ouvinte. Mas a influência atua nos dois sentidos: através da escolha que fazem, os falantes mostram sua compreensão do contexto – suas escolhas linguísticas simultaneamente criam o contexto (id, p. 6-7).

Mais uma vez, lembramos que a premissa básica é de que toda língua pode ser vista como conjuntos de escolhas em contexto. Com relação à Linguística de Corpus, três observações bastante básicas podem ser feitas acerca dos resultados já obtidos através de abordagens da descrição linguística baseadas em corpus, as quais têm como ponto de partida o comportamento recorrente, como observado no corpus, de itens lexicais individuais (Thompson and Hunston 2007, p. 9-10). A primeira é que itens lexicais são usados de forma surpreendentemente regular; regularidade essa que pode ser calculada através da colocação¹. A segunda, que é consequência dessa regularidade, é que pode-se ver muitos itens lexicais ocorrendo em frases semifixas e/ou co-textos semânticos ou discursivos específicos; o termo ‘fraseologia’ pode ser usado para descrever a tendência de algumas palavras de ocorrer não de forma aleatória e apenas baseada na gramática, mas sim em sequências preferenciais (Hunston 2002, p. 137-138). E por fim, é observado que o significado e a forma estão ligados intimamente, no sentido de que se uma palavra tem mais de um significado, pelo menos um desses significados deve tender a ser associado com uma frase ou sequência em particular. Essas observações levaram à concepção de duas teorias (id, p. 10-11). A primeira afirma que o significado não está nas palavras individuais, mas em unidades de significado de extensão indeterminada. A segunda teoria é de que a linguagem não existe como um sistema gramatical no qual os itens lexicais ocupem espaços definidos, mas como uma série de frases semifixas. Se essas teorias estão corretas, elas põem em dúvida uma visão de linguagem segundo a qual as escolhas gramaticais são em um certo sentido anteriores às escolhas lexicais, e que vê o item lexical individual como o nível final de sistemas de

¹ Colocação: palavras tendem a ocorrer seguindo padrões, mais ou menos frequentemente junto com outras, com ou sem posições cristalizadas (p. ex., “completamente diferente” é mais provável que “completamente igual”) (cf. Hunston 2002, p. 138).

especificidade crescente. Ou seja, confirma a noção de que há uma junção entre léxico e gramática, conforme também proposto pela LSF.

A Linguística de Corpus envolve a investigação de grandes quantidades de texto naturalmente produzido, que são armazenadas eletronicamente. Segundo Thompson e Hunston (id, p. 8), muito do trabalho feito nessa área com anotação de corpus assume que o papel de um corpus é oferecer evidência quantitativa para diversas abordagens da linguagem. E o trabalho quantitativo em LSF é consistente com esta visão de corpora.

A teoria da LSF pressupõe uma gramática complexa, em que existem três análises possíveis para cada oração; cada uma dessas análises utiliza grupos de redes sistêmicas separadas, embora interdependentes, e cada aspecto da descrição das características da linguagem é enriquecido por sua conexão com uma teoria semiótica mais geral (id, p. 2). A Linguística de Corpus, por outro lado, é “quase perversamente ‘leve de teoria’ e leva a descrições menos detalhadas [“thin descriptions”]” (id, p. 2). Para muitos que adotam uma abordagem baseada em corpus, a abordagem inicial dos dados deveria ser “neutra” em termos de teoria – embora na prática, como lembram Thompson e Hunston (2007, p. 2), não sejam tão neutros assim, por estarem fundamentados no pressuposto de que o comportamento recorrente de itens lexicais individuais é linguisticamente significativo.

Como apresentado no Capítulo 3, a LSF vê a linguagem como potencial de significado, que pode ser representado como um conjunto de redes sistêmicas, cada uma mapeando uma série sucessiva de possíveis realizações de significados. Já a Linguística de Corpus vê a totalidade do uso da língua, parcialmente observável em um corpus. Assim, o conceito de potencial de significado é substituído pelo conceito de ocorrência: o foco é menos no que poderia acontecer e mais no que realmente acontece. O significado reside na recorrência em um grande número de textos (Thompson e Hunston 2007, p. 2). Além disso, numa visão extremada, pode-se dizer que a LSF considera que as questões mais essenciais sobre a linguagem já foram respondidas pela teoria existente, e novas observações devem ser incorporadas a essa visão ou descartadas; enquanto que a Linguística de Corpus considera que muitas questões sobre a linguagem ainda estão em aberto, e insiste em ver os detalhes como importantes (id, p. 3-4)

4.2 O trabalho baseado em Corpus

Como ressalta Kennedy (1998, p. 2), a Linguística de Corpus não começou com o desenvolvimento dos computadores, mas não há dúvida de que ela foi enormemente beneficiada por esse progresso, que tanto facilitou a análise baseada em texto quanto permitiu a ampliação do tamanho das bases de dados usadas para análise – o corpus Bank of English, por exemplo, passa de 320 milhões de palavras (cf. Sardinha 2000, p. 331).

Segundo Hunston (2002, p. 38), o uso do *Concordancer* é talvez a forma mais básica de processar a informação do corpus, e a maioria das pesquisas na área se volta para as linhas de concordância e sua interpretação. O *Concordancer* é um programa que faz buscas de palavras ou expressões no corpus e apresenta as ocorrências encontradas com um pequeno contexto de cada lado, como pode ser visto na figura abaixo.

... ferida entrevista." Na mesma ocasião, o jornalista de 68 anos pediu desculpas ao público a ...
 ... m nós", disse ele. Fino foi o primeiro jornalista de todo o mundo a noticiar o começo da ...
 ... UA têm estratégia errada', diz Arnett Jornalista deu entrevista à Folha antes de ter sid ...
 ... alista japonês explodiu. Interrogado, o jornalista disse que trazia a peça do Iraque como ...
 ... jetivo de impedir um êxodo de civis. Um jornalista do diário americano "The New York Times ...
 ... osta, na Bela Vista. "Ensaio Geral - O Jornalista e Todas as Guerras" enfeixa textos e pe ...
 ... m ato panfletário." ENSAIO GERAL - O JORNALISTA E TODAS AS GUERRAS. Quando: hoje, às 21 ...
 ... no fatura no mínimo US\$ 500 por dia por jornalista. * Eles só aceitam pagamento em dinh ...
 ... ão?". Ninguém riu. * É cara a vida do jornalista em Bagdá em tempos de guerra. Gastam-se ...
 ... Michael Kelly, 46, tornou-se o primeiro jornalista "embutido" na cobertura montada pelo Pe ...
 ... rtualidade que se incorpora às tropas o jornalista "encaixado" [dotado de câmara e equipam ...
 ... australiano e dois britânicos. Um quinto jornalista está desaparecido. 03/04/2003 Jorna ...
 ... ava entre US\$ 250 e US\$ 750 por dia por jornalista estrangeiro para permitir que este trab ...
 ... vas que caíram a centenas de metros. Um jornalista europeu se surpreendeu: "Ainda tem peri ...
 ... ue não era um fuzil, não era", disse um jornalista europeu, cinco conflitos nas costas. * ...
 ... DIAS DE NOVA YORK Para o escritor e jornalista Gay Talese, o Pentágono alcançou o seu ...
 ... sse. Antes, durante o fim de semana, o jornalista havia dado uma entrevista à Folha em Ba ...
 ... sta cai do telhado e morre no Iraque O jornalista húngaro Gaby Rado, 48, morreu ontem ao ...
 ... ue e libertar o povo iraquiano. Como o jornalista insistisse que "desarmar" significava i ...
 ... Agora, George?", texto do dramaturgo e jornalista Izaías Almada. Gôes investe em criação ...
 ... efato de metal que estava na mala de um jornalista japonês explodiu. Interrogado, o jornal ...
 ... EUA no hotel Palestine TAREQ AYYOUB, jornalista jordaniano da rede de TV Al Jazeera, do ...
 ... possível consegui-los através do e-mail jornalista.jorge@uol.com.br ou pelo tel. 3259-8408 ...
 ... ue afirmam que muito pior do que ter um jornalista junto às tropas seria não ter nenhum. ...
 ... foi capturado, e eis o quadro geral do jornalista médio em Bagdá. * Ontem de manhã, enq ...
 ... agências internacionais 15/04/2003 Jornalista morre em viagem a Bagdá O jornalista a ...
 ... elefone por satélite. 23/03/2003 Um jornalista morre, e dois desaparecem DA REDAÇÃO ...
 ... três combatentes curdos. Paul Moran, o jornalista morto, era cinegrafista free-lance do c ...

Figura 5 – Concordâncias extraídas do corpus deste trabalho

Uma aplicação da pesquisa baseada em corpus pode ser a análise da fraseologia (Hunston 2002). A autora considera como parte dessa análise a questão de frases como “onde há fumaça há fogo”, que podem ser exploradas em construções recorrentes, mas livres ou adaptadas, como “o povo brasileiro está começando a desconfiar de que onde há fumaça, há mesmo fogo” ou “o escândalo mais recente mostrou que às vezes pode haver fumaça sem haver fogo”; ou a tendência de certos verbos de ocorrerem na voz passiva e não na ativa, ou na forma negativa e não na positiva, ou no presente ou no passado, por exemplo, “nunca me passou pela cabeça fazer isso” parece mais provável do que “sempre me passou pela cabeça fazer isso” ou “nunca me passará pela cabeça fazer isso”. A fraseologia, diz a autora, é mais que uma coleção de frases: abrange todos os aspectos de sequenciamentos preferidos, bem como a ocorrência das chamadas frases fixas (Hunston 2002, p. 138).

Para Hoey (2007, p. 53), por consideramos que a linguagem é feita de escolhas, precisamos reconhecer que o que é escolhido é o item lexical com suas colocações, suas coligações² e suas associações semânticas. Stubbs (2007, p. 17) considera que as observações mais importantes da Linguística de Corpus estão justamente relacionadas às descobertas sobre coocorrências, evidenciadas através de concordâncias que demonstram padrões repetidos e regularidades. Segundo ele, a importância dessas observações se deve ao fato de que demonstram um aspecto do uso da linguagem subestimado em descrições prévias: a disseminação de combinações semifixas com padrões lexicogramaticais e semânticos definidos. “É a descoberta dessas unidades funcionais de significado que tem as implicações mais importantes para a teoria e a descrição linguísticas, já que não correspondem a unidades lexicais ou gramaticais tradicionais. Em outras palavras, há camadas de organização em uma língua das quais anteriormente havia apenas, no máximo, ligeiras suspeitas” (Stubbs 2007, p. 17). A análise através de programas de computador, portanto, pode contribuir para descobertas linguísticas, pois evidências que emergem dos dados podem ressaltar como relevantes aspectos antes não observados – ou observáveis (Oliveira 2009).

² Coligações são as “companhias gramaticais” em que o item aparece, ou não aparece, tanto no seu grupo ou em um grupo hierarquicamente superior, bem como as funções gramaticais que esse grupo prefere ou evita – cf. Hoey 2007, p. 38.

Também a expressão de ideologias e atitudes já foi investigada com o uso de técnicas baseadas em corpus, lembra Conrad (2002, p. 80), especialmente com relação às associações entre palavras. Por exemplo, Coffin e O'Halloran (2006) demonstram, através da comparação de um corpus especializado – textos do tablóide *The Sun* – com o corpus do Bank of English, como o uso recorrente de certas palavras em certos contextos confere a essas palavras conotações ideológicas implícitas. Conrad (2002, p. 78) explica que reconhecer padrões no uso da linguagem requer, necessariamente, verificar se um fenômeno é comum ou incomum – ou seja, fazer uma verificação quantitativa. Mas, ressalta ela, ao mesmo tempo, os números por si só fornecem pouca informação sobre a linguagem. Frequentemente, como lembra Oliveira (2009, p. 3), para interpretar os dados, com base no corpus, é preciso também levar em conta o co-texto e os aspectos socioculturais ligados aos textos. Dados por si só oferecem evidências linguísticas; cabe ao pesquisador analisá-los e interpretá-los para assim extrair deles informação (Hunston 2002, p. ; Oliveira 2007, p. 57).

Stubbs (2007, p. 28) ressalta que muito da análise de corpus que tem sido feita tem apresentado objetivos estritamente linguísticos, além de um cunho muito mais descritivo que explicativo. Falta agora, segundo ele, um movimento de tentar relacionar essas descobertas às “as grandes questões da teoria linguística, cognitiva e social” (id, p. 34). Ou seja, buscar uma forma de fazer pesquisa com corpus que leve também a investigar implicações socioculturais. Stubbs afirma que os estudos de corpus têm começado a mostrar que há muitos casos em que significados pragmáticos estão associados convencionalmente a formas linguísticas. Teóricos da fraseologia têm apontado que expressões idiomáticas são usadas com frequência para codificar uma ‘situação recorrente de interesse social específico’ (id, p. 27). Ele sugere, como possibilidade de unir a descrição à explicação, o estudo de áreas de significado em que a linguagem utiliza fraseados recorrentes para expressar conhecimento cultural e em que existe categorização significativa (id, p. 32-33).

Kennedy (1998, p. 61) explica que as questões relacionadas a design e construção do corpus são basicamente ligadas à validade e confiabilidade da pesquisa baseada naquele corpus, e incluem decidir se o corpus será estático ou dinâmico, como proceder para que ele seja o mais representativo possível de uma

língua ou um gênero, quão extenso ele deve ser, e que tamanho devem ter as amostras de textos.

No Brasil, a pesquisa em Linguística de Corpus está em crescimento. Projetos de compilação de corpus estão sendo desenvolvidos, como da PUC-SP, especializado em textos de comunicação no contexto de negócios, ou o corpus do Núcleo Inter-Institucional de Linguística Computacional – NILC, da USP-São Carlos/UFSCar/Unesp (Oliveira 2009). Na PUC-Rio, o projeto CORPOBRAS, iniciado em 2002, busca montar um corpus representativo do português do Brasil. Em 2008, o CORPOBRAS ultrapassou a meta de um milhão de palavras. Diversos trabalhos foram e estão sendo feitos com base nesse corpus, como Caldeira (2006), que estudou redações de vestibular, observando, por exemplo, que nominalizações foram menos frequentes que processos nestes textos, o que as diferencia, por exemplo, de artigos acadêmicos, onde as nominalizações são muito frequentes e variam de acordo com as áreas de conhecimento (Oliveira, 2006).

Este trabalho, por suas características estarem em conformidade com as questões discutidas acima, se insere na pesquisa em Linguística de Corpus. A elaboração do corpus utilizado e as ferramentas de análise aplicadas serão explicadas a seguir, no Capítulo 5.