4
Interfaces: A abordagem de corpus e a sistêmico-funcional

Um corpus não contém novas informações sobre a língua, mas o programa computacional nos oferece uma nova perspectiva sobre o que já é familiar.[1]

(Huston, 2002:3)

A Linguística de Corpus (doravante LC) tem sido considerada por muitos como a face moderna da linguística empírica (Teubert, 1996). Além de caracterizar uma nova forma de estudo e concepção da linguagem, a LC tem uma maneira específica de fazer análises linguísticas e está baseada no estudo de textos reais com o auxílio de programas computacionais visando a extração de evidências linguísticas. Um corpus linguístico de base computacional, portanto, é concebido como coleções de textos que ocorrem naturalmente na língua, organizadas sistematicamente para representar áreas de uso da língua, e das quais podemos extrair novas informações (Biber, 1995, p.310).

Historicamente, a Linguística de Corpus é uma área de estudos recente, principalmente, se comparada a outras correntes teóricas de linguagem. Contudo, embora tenha menos de 50 anos de desenvolvimento, a LC é uma das áreas da linguística que mais tem se desenvolvido nos últimos anos. Vários fatores contribuíram para tal desenvolvimento, dentre eles: maior acesso ao computador, sofisticação de programas computacionais para análise linguística, aumento no número de pesquisadores e associações ligadas à LC, desenvolvimento de diferentes abordagens de ensino que têm um ponto de interseção com a LC etc (Sardinha, 2000). No Brasil, embora a área ainda seja relativamente recente, é possível perceber que cada vez mais pesquisadores, professores e alunos começam a utilizar a LC como um suporte não apenas teórico, mas também aplicado ao ensino, o que tem contribuído para a criação de materiais que facilitam o processo de aprendizagem (Vianna & Tagnin, 2010).

[1] Tradução livre do original: "A corpus does not contain new information about language, but the software offers us a new perspective on the familiar."

Atualmente, vários estudos teóricos e aplicados ao português do Brasil têm sido desenvolvidos com base no uso de *corpora* para a descrição de fenômenos linguísticos ou para a verificação de hipóteses acerca dos mesmos[2]. Como este estudo, muitas dessas pesquisas têm sido realizadas em interface com a Linguística Sistêmico-Funcional, visando a investigação linguística de gêneros do contexto escolar (Castro, 2009; Nóbrega, 2009; Ramos, 2010), já que a LC é uma área situada na interdisciplinaridade e na complementaridade (Oliveira, 2009, p. 52).

Seguindo essa tendência, neste estudo, a LC foi utilizada em interface com a LSF a partir de duas perspectivas: tanto como uma abordagem baseada em corpus (*corpus-based approach*), já que ela foi utilizada (1) para ampliar, descrever e confirmar, através de resultados quantitativos, o uso de um fenômeno linguístico específico, a função coesiva das nominalizações, em um gênero textual; (2) quanto como uma abordagem orientada pelo corpus (*corpus-driven approach*), pois as evidências linguísticas identificadas no corpus averiguado possibilitaram *novas proposições descritivas com consequências teóricas* (Oliveira, 2009, p.60) do fenômeno investigado. Em outras palavras, a LC funcionou como um suporte para a busca de evidências para o desenvolvimento de descrições mais específicas sobre o funcionamento das nominalizações como elo coesivo nas redações escolares, que é um assunto pouco explorado na literatura vigente.

Os pressupostos básicos que caracterizam a LC têm norteado o desenvolvimento da área e são compartilhados com outras correntes teóricas, principalmente as de cunho funcionalista, em especial, a Linguística Sistêmico-Funcional, que é a teoria linguística utilizada para a realização deste estudo, conforme serão brevemente discutidos a seguir.

✓ A língua como um fenômeno social

Um dos pressupostos que caracteriza a LC como uma vertente que tem um olhar *diferente sobre a linguagem* (Teubert e Čermáková, 2007, p.35) está relacionado à concepção de que a língua é um fenômeno social, o que difere significantemente das concepções mais formalistas da linguagem, em que o conceito de língua é concebido quanto ao seu aspecto formal, isto é, em relação às propriedades internas de suas

[2] Oliveira (2009, p.63-67) apresenta alguns desses trabalhos que visam a descrição de diversos usos do português.

formas, e a mesma é vista, portanto, como um sistema de caráter abstrato e autônomo que tem como principal função a expressão de pensamento.

Uma vez que a língua é concebida como um fenômeno social, à LC interessa investigar como significados são construídos, e, para tanto, ela os concebe como aquilo que pode ser verbalmente comunicado e socialmente estabelecido por membros de uma comunidade discursiva. Dessa forma, um dos interesses genuínos e principais desta vertente linguística é verificar de que forma palavras, sentenças e textos ganham significados em função dos diferentes cotextos em que se encontram. Para tanto, a linguagem é investigada através do discurso, que está relacionado à própria concepção da palavra *corpus:*

> Ainda que nos limitemos a textos que foram preservados, esse discurso é muito grande para ser investigado. Nunca será possível estudar todos os textos que existem. Tudo o que a linguística de corpus pode fazer é trabalhar com uma amostra (apropriada) do discurso. Essa amostra é chamada *corpus* (Teubert e Čermáková, 2007 p.41)[3].

Em termos teóricos, dessa maneira, é possível conceber uma complementariedade entre a Linguística de Corpus e a abordagem sistêmico-funcional, já que ambas as vertentes concebem a linguagem enquanto fenômeno social, em função de seu uso e sua funcionalidade no discurso.

✓ Dados autênticos

Um dos principais aspectos que caracteriza a LC como tal é a origem dos dados. Isso significa dizer que, além de o conteúdo de um corpus ter que ser criteriosamente escolhido, a fim de garantir a representatividade do mesmo face aos objetivos da pesquisa que se quer desenvolver, os dados devem necessariamente ser autênticos, isto é, devem representar a língua tal como ela é utilizada por seus falantes em diferentes contextos sociais. Tal pressuposto tem implicações diretas não só quanto à concepção do que seria a língua, mas também na relação entre ensino/aprendizagem, uma vez que o foco do processo deixa de ser exclusivamente a variante culta, padrão, da língua e

[3] Tradução livre do original: 'Even if we confine ourselves to the texts that have been preserved, this discourse is much large to ma.ke it (…). It will never possible to study all extant texts. All corpus linguistics can do is to work with a (suitable) sample of the discourse. Such a sample is called *corpus'* (Teubert e Čermáková, 2007 p.41)

passa a ser a variante do uso, aquela que o indivíduo encontra e usa em diferentes esferas da comunidade linguística em que está inserido.

Dessa forma, a vantagem de se ter um banco de dados com exemplos linguísticos reais influencia diretamente em uma mudança de paradigma linguístico, que é caracterizado por Huston da seguinte forma:

> Um corpus não contém novas informações sobre a língua, mas o programa computacional nos oferece uma nova perspectiva sobre o que já é familiar. (Huston, 2002 p.3)[4]

Na citação acima, também em epígrafe neste capítulo, a autora explica que o ponto em questão não é que o corpus traga novas informações sobre a língua; ele possibilita uma nova perspectiva sobre a mesma, de caráter empírico, que analisa padrões reais de uso e que acaba por romper com a perspectiva racionalista quanto à linguagem, vigente e marcada especialmente por trabalhos como os do linguista Noam Chomsky (Sardinha, 2000).

Assim, a autenticidade dos textos é outro ponto de aproximação entre a abordagem de corpus e a vertente sistêmico-funcional, já que em ambas as áreas as análises linguísticas são realizadas em textos reais, isto é, textos que ocorrem naturalmente nos contextos sociais em que a língua é usada.

✓ Padrões léxico-gramaticais

Um padrão, segundo Sinclair (1991), é definido como 'uma sequência recorrente de (pelo menos duas) palavras, dentro de um espaço delimitado (normalmente equivalente a até quatro palavras de distância) que possui um sentido específico'[5]. Os padrões são, de acordo com Sardinha (2006), 'um tipo de unidade pré-fabricada da língua que parece estar disponível por inteiro na memória do indivíduo tanto para a produção quanto para a recepção linguística'. O padrão léxico-gramatical, portanto, caracteriza-se justamente pelo fato de ele residir tanto no nível lexical quanto no nível gramatical.

[4] Tradução livre do original: "A corpus does not contain new information about language, but the software offers us a new perspective on the familiar" (Huston, 2002, p.3).

[5] O termo *padrão* pode ser entendido como o que é definido por *colocação* (*collocation*), já que esta é definida como uma co-ocorrência significativa e habitual de duas ou mais palavras em proximidade uma das outras (Teubert e Čermáková, 2007, p. 139).

A partir desses padrões, concebe-se, mais uma vez, uma mudança de perspectiva quanto à concepção de linguagem, já que, a partir deles é possível falar sobre uma natureza associativa e probabilística da linguagem: associativa porque diz respeito a estruturas linguísticas a partir de agrupamentos padronizados de palavras e não em unidades isoladas; e probabilística porque certos traços – padrões – são mais frequentes do que outros (Halliday, 1991, 1992 *apud* Sardinha 2000).

As noções de associação e de probabilidade estão presentes tanto nos pressupostos teóricos da Linguística de Corpus, que adota o conceito de *colocação*, o qual pode ser entendido como probabilidades de uso de algumas palavras com outras, quanto nos da Linguística Sistêmico-Funcional, que entende a língua como um *potencial de significados*, isto é, como possibilidades de escolhas, as quais podem ser vistas como probabilidades de usos de diferentes elementos do sistema linguístico em determinados contextos culturais ou situacionais, (cf. capítulo 2, item 2.2, p. 25) o que mostra uma consonância entre essas duas vertentes.

Por fim, vale ressaltar a importância do uso do computador e de ferramentas de processamento linguístico na Linguística de Corpus, tanto para a identificação dos padrões quanto para análise dos mesmos. Embora em termos metodológicos, a vertente sistêmico-funcional não seja caracterizada pelo uso de programas computacionais para a análise linguística, muitos estudos com base nessa teoria já têm sido realizados com a ajuda de ferramentas computacionais para investigações linguísticas sobre a ocorrência de padrões léxico-gramaticais (Oliveira, 2006, Valério et al, 2007, Valério & Oliveira, 2009, dentre outros). Dessa maneira, nos estudos supracitados, e neste trabalho, enquanto essas frequências são medidas em termos quantitativos pela LC, elas são interpretadas qualitativamente com base nos pressupostos da LSF, o que também reflete uma complementariedade entre essas duas vertentes.

Assim sendo, a partir dos pressupostos básicos mencionados acima, a LC é uma área do saber multidisciplinar, que se agrupa a outras várias áreas de conhecimento e teorias. Logo, a aproximação dessa vertente com a perspectiva sistêmico funcional da linguagem é evidente, uma vez que ambas priorizam o aspecto social e funcional da linguagem, a partir da análise de textos reais, ocorridos naturalmente na língua, a fim de investigar as 'probabilidades de colocação [escolha] de palavras com outras em determinados contextos de uso da língua' (Oliveira, 2009, p.53).

Além disso, por conceberem e utilizarem o conceito de gênero discursivo de uma perspectiva mais abrangente, considerando-o em função dos diferentes usos da

língua (Swales, 2002 apud Sardinha 2000), também é possível assumir uma convergência entre a LC e os pressupostos da LSF, que define os gêneros como processos sociais que são realizados com um objetivo comunicativo em uma dada cultura (cf. capítulo 2, seção 2.2, p. 26).

No que concerne à adoção da LC como aparato teórico, ela tem sido considerada por muitos apenas como uma metodologia, e, como qualquer metodologia, a mesma é suscetível a uma desatualização. Contudo, como já mencionado aqui, a LC é uma área de conhecimento em expansão, que apresenta sua própria maneira de olhar a linguagem e formas próprias e empíricas de manipulação de dados. Além disso, essa área apresenta variadas metodologias para análise linguística, tais como buscas em contexto (Scott, 2004) e a análise multidimensional (Oliveira, 1997, Biber, 1988). Dessa forma, parece estar acontecendo uma mudança de paradigma advinda do posicionamento de alguns pesquisadores quanto à concepção teórica da LC: muitos estão começando a considerá-la como uma área de conhecimento propriamente dita, e não apenas como uma metodologia ou um mero instrumento de suporte para outras teorias linguísticas[6]. Nesse sentido, essa alteração de perspectiva está relacionada ao fato de a LC investigar e conceber, de uma perspectiva empirista, o objeto com o qual pesquisadores, professores e alunos, têm contato diário: a língua.

Acredita-se, contudo, que a adoção de ambas as abordagens caracteriza um importante passo em relação a uma mudança de perspectiva investigativa e didática, não apenas no que concerne o aspecto coesivo das nominalizações, mas também aos estudos da linguagem como um todo. Dessa forma, tomar a LC como um aparato teórico significa considerá-la como uma fonte inesgotável quanto à descrição e à análise de fenômenos linguísticos. Nesse sentido, é necessário ressaltar o grande potencial da LC para fins de análise linguística, a partir de um viés teórico e descritivo sobre a língua, que valoriza a língua em uso. Para que tal posicionamento seja possível, é necessário adotar a concepção sugerida por Oliveira, a qual foi adotada neste estudo, de que é:

> (...) necessário deixar de pensar que a Linguística de Corpus se restringe à compilação e coleta de dados, já que ao contribuir para a geração de novas descrições das línguas ela contribui também para que possamos conhecer novas gramáticas, que por sua vez nos levam a entender melhor a experiência humana tal como é construída na linguagem. (Oliveira, idem, p.67)

[6] Para uma maior e mais detalhada discussão sobre este assunto ver Oliveira, 2009.

No próximo capítulo serão apresentados de que forma os recursos da LC, junto aos fundamentos teóricos da LSF, foram utilizados na realização deste trabalho, mais especificamente, na análise dos dados selecionados para este estudo, caracterizando os pressupostos metodológicos desta pesquisa.