2
Lingüística de corpus

Present real examples only
This is perhaps the first lesson to be learned from corpus study. Language cannot be invented; it can only be captured.[4]

(Sinclair, 1997, p. 31)

Este primeiro capítulo dedicado à discussão dos pressupostos teóricos da pesquisa privilegia os aspectos mais relevantes da lingüística de corpus para o estudo em tela. Assim sendo, enfoca-se inicialmente a noção de corpus, tanto em termos de sua definição como de sua caracterização. Os padrões de uso de uma língua, que são trazidos para a agenda de pesquisa pela lingüística de corpus, também são detidamente considerados neste capítulo. Por fim, oferece-se um resumo dos tópicos abordados.

2.1.
Corpus

Com o desenvolvimento da lingüística de corpus, várias definições têm sido propostas para o termo 'corpus'. Apesar de haver concordância a respeito de certos elementos definidores, algumas diferenças são observadas nas propostas de autores distintos.[5]

---

[4] Tradução livre para o português: "Apresente somente exemplos reais // Esta é talvez a primeira lição a ser aprendida a partir do estudo de corpus. A linguagem não pode ser inventada; ela só pode ser capturada".

[5] A análise das definições realizada a seguir baseia-se principalmente no que é expresso pelos autores e não no que os mesmos possivelmente deixam de incluir nestas.

A definição[6] oferecida por Tribble & Jones (1990, p. 8) em seu livro inaugural acerca do uso de linhas de concordância na sala de aula indica que: "*large compilations of documents by a particular author or representative of a particular genre, or else writings on a particular topic. Such a compilation of texts is called a corpus (plural corpora)*".[7] A explicação esclarece um primeiro elemento na constituição de um corpus, que é a sua representatividade. No entanto, os autores afirmam ser este uma coletânea de documentos sem especificar o que seria entendido por este termo. Ao indicar o que um corpus pode representar, mencionam-se autores, gêneros ou tópicos, mas não é feita nenhuma referência a línguas. Em relação ao tamanho, apesar de os corpora serem grandes, especialmente devido ao aumento da capacidade de processamento dos computadores, não parece esta ser uma característica definidora dos mesmos.

Em sua obra seminal, Sinclair (1991, p. 171) define o termo da seguinte forma:

> *A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are the clues to the lexical structure of the language.*[8]

Nesta proposta, Sinclair (1991) traz à discussão os conceitos de 'coleção' e 'caracterização' apesar de não empregar exatamente estes termos. A noção dúbia de documentos (cf. Tribble & Jones, 1990) é substituída por textos e a função do corpus como representativo de uma língua é adicionada. Além disto, apesar de mencionar a questão do tamanho, o autor opta por não incluí-la na definição de corpus, mas em sua caracterização típica.

---

[6] Diferentemente do padrão adotado nesta dissertação, inverte-se a ordem de apresentação nesta seção. No corpo do texto, são indicadas as definições originais em inglês de cada autor analisado de forma que sejam mantidas as nuances de significado por eles propostas. A tradução para o português é oferecida em notas de rodapé.

[7] Tradução livre para o português: "grandes compilações de documentos de um autor particular ou representativas de um gênero específico, ou então escritos sobre um mesmo tópico. Tal compilação de textos é chamado de um corpus (plural corpora)".

[8] Tradução livre para o português: "Um corpus é uma coleção de textos que ocorrem naturalmente, escolhidos para caracterizar um estado ou uma variedade de uma língua. Na lingüística computacional moderna, um corpus tipicamente contém muitos milhões de palavras: isto é porque se reconhece que a criatividade da linguagem natural leva a uma imensa variedade de expressão que é difícil isolar padrões recorrentes que são indícios da estrutura lexical da língua".

Por sua vez, Aarts (1991, p. 45) opta por uma definição mais concisa: "*a corpus is understood to be a collection of samples of running text. The texts may be in spoken, written or intermediate forms, and the samples may be of any length*".[9] A importância desta definição é a inclusão da possibilidade de o corpus conter amostras de textos e nem sempre textos completos. No entanto, quando se trata de tamanho é preciso notar que se as amostras forem de tamanhos altamente distintos pode ser que o corpus passe não mais a representar a totalidade da população investigada, mas seja veículo de expressões idiossincráticas. Este problema parece inexistir quando o objetivo é justamente o levantamento de usos particulares como ocorre, por exemplo, na investigação do estilo empregado por um determinado autor com o auxílio de um corpus que contemple única e exclusivamente todas as suas obras literárias.

A definição de corpus oferecida no manual do *MicroConcord*, um programa de análise textual da *Oxford University Press*, revela-se simples: "*Technically, the word you look for is known in* MicroConcord *as the SEARCH WORD and the texts you look in constitutes the CORPUS*"[10] (Murison-Bowie, 1993, p. 8). Mesmo assim, a novidade que se apresenta de forma indireta por Murison-Bowie (1993) é que o corpus é investigado por meio de uma ferramenta computacional. Mais tarde, indica-se ser o corpus um sinônimo para "quantidades de textos" (Murison-Bowie, 1993, p. 49). Ambas as propostas são modestas, mas isto é possivelmente intencional de forma a permitir a compreensão do programa por parte de seus usuários.

Em um livro dedicado à lingüística de corpus, McEnery & Wilson (1996, p. 21) argumentam que

> *the term 'corpus' when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings: sampling and representativeness, finite size, machine-readable form, a standard reference.*[11]

---

[9] Tradução livre para o português: "um corpus é entendido como uma coleção de amostras de textos corridos. Estes textos podem estar na forma oral, escrita ou intermediária, e as amostras podem ser de qualquer tamanho".

[10] Tradução livre para o português: "Tecnicamente, a palavra que você procura é conhecida no *MicroConcord* como a PALAVRA DE BUSCA e os textos que você examina constituem o CORPUS".

[11] Tradução livre para o português: "o termo 'corpus' quando usado no contexto da lingüística moderna tende mais freqüentemente a ter conotações mais específicas do que esta simples definição provê. Estas podem ser consideradas à luz de quatro tópicos principais:

Neste trecho, os autores explicitam os princípios que guiam a criação de um corpus. Esta proposta introduz a noção de que um corpus tem um tamanho finito apesar de a língua representada no mesmo ser infinita. Outra diferença desta proposta é a de considerar o corpus como uma "referência padrão".

Aston (1997, p. 205) adota uma definição sintética para o termo: "*computer readable collection of texts or transcripts which can be accessed and interrogated selectively using text retrieval or concordancing software*".[12] A inovação desta proposta é justapor duas referências à ferramenta computacional tanto no formato do corpus quanto no processamento do mesmo.

Outra definição sucinta é oferecida por Kennedy (1998, p. 1): "*In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description*".[13] A contribuição desta definição é a de registrar explicitamente o objetivo de um corpus, isto é, análise e descrição lingüísticas.

Em *Longman grammar of spoken and written English*, os autores oferecem a seguinte definição de corpus: "*a collection of spoken and written texts, organized by register and coded for other discourse considerations, comprises a corpus*"[14] (Biber et al., 1999, p. 4). A explicação é resumida talvez dado que o escopo da mesma não é explicar os conceitos da lingüística de corpus. No entanto, a indicação de que um corpus é "codificado para outras considerações discursivas" pode ser verdade no caso do corpus empregado para o desenvolvimento da gramática, mas tipo de anotação não é característico de todo e qualquer corpus.

A partir de todo um desenvolvimento anterior, a proposta de Tognini-Boneli (2001, p. 2) aponta para uma definição mais ampla:

---

amostragem e representatividade, tamanho finito, forma que pode ser lida por máquina e uma referência padrão".

[12] Tradução livre para o português: "coleções de textos ou transcrições passíveis de serem lidas pelo computador que podem ser acessadas e examinadas seletivamente usando programas de busca textual ou concordanciadores".

[13] Tradução livre para o português: "Nas ciências da linguagem, um corpus é um corpo de texto escrito ou fala transcrita que pode servir de base para a análise e descrição lingüística".

[14] Tradução livre para o português: "uma coleção de textos orais e escritos, organizada por registro e codificada para outras considerações discursivas, abrange um corpus".

> *A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. Usually the assumption is that the language stored in a corpus is naturally-occurring, that it is gathered according to explicit design criteria, with a specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology. Not everybody, of course, goes along with these assumptions but in general there is consensus that a corpus deals with natural, authentic language.*[15]

Aqui a autora explicita a necessidade de haver um propósito determinado para a compilação do corpus. Este propósito deve anteceder tal compilação e deve iluminar os critérios de seleção acerca do que incluir no corpus.

As duas definições seguintes reiteram as noções discutidas anteriormente em relação ao significado do termo corpus:

> *A corpus is a large, principled collection of naturally-occuring texts that is stored in electronic form (accessible on computer). Corpora can include both written and transcribed spoken texts.*[16] (Conrad, 2002, p. 76)

> *A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria. There are four important characteristics to note here: 'authentic', 'electronic', 'large' and 'specific criteria'.*[17] (Bowker & Pearson, 2002, p. 9).

Uma mesma restrição pode ser feita às duas propostas: a ênfase dada ao tamanho do corpus, quando este aspecto não é um elemento essencial para a identificação de uma coleção de textos como um corpus.

Hunston (2002, p. 2), diferentemente das definições anteriormente revisadas, detalha melhor o conceito:

---

[15] Tradução livre para o português: "Um corpus pode ser definido como uma coleção de textos que se assume ser representativa de uma dada língua, compilada de forma que possa ser usada para análise lingüística. Usualmente a pressuposição é que a língua armazenada em um corpus ocorre de forma natural, que ela é colhida de acordo com critérios de planejamento explícitos, com um propósito específico em mente, e com uma reivindicação de representar padrões maiores da língua selecionados de acordo com uma tipologia específica. Nem todo mundo, evidentemente, aceita estas pressuposições, mas em geral há um consenso que um corpus contempla língua natural, autêntica".

[16] Tradução livre para o português: "Um corpus é uma grande e criteriosa coleção de textos que ocorrem naturalmente que é armazenada em formato eletrônico (acessível no computador). Os corpora podem incluir tanto textos escritos como transcrições de textos orais".

[17] Tradução livre para o português: "Um corpus pode ser descrito como uma grande coleção de textos autênticos que foram reunidos em formato eletrônico de acordo com um conjunto específico de critérios. Há quatro características importantes a serem notadas aqui: 'autêntico', 'eletrônico', 'grande' e 'critérios específicos'".

> *A corpus is defined in terms of both its form and its purpose. Linguists have always used the word corpus to describe a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study. More recently, the word has been reserved for collections of texts (or parts of text) that are stored and accessed electronically. Because computers can hold and process large amounts of information, electronic corpora are usually larger than the small, paper-based collections previously used to study aspects of language. A corpus is planned, though chance may play a part in the text collection, and it is designed for some linguistic purpose. The specific purpose of the design determines the selection of texts, and the aim is other than to preserve the texts themselves because they have intrinsic value. This differentiates a corpus from a library or an electronic archive. The corpus is stored in such a way that it can be studied non-linearly, and both quantitatively and qualitatively. The purpose is not simply to access the texts in order to read them, which again distinguishes the corpus from the library and the archive.*[18]

Vários aspectos são ressaltados na definição / explicação como, por exemplo, a diferença entre corpora por um lado, e bibliotecas e arquivos eletrônicos por outro, além da variação dos primeiros ao longo do tempo. Neste sentido, esclarece-se o uso do termo "maior": ele é empregado em comparação às iniciativas pré-computacionais.

Uma nova definição de corpus é proposta por Sinclair (2005) após 14 anos da publicação de *Corpus concordance collocation* (Sinclair, 1991): "*a corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research*".[19] Se esta definição for comparada à de 1991 várias diferenças podem ser notadas. O corpus passa a ser, para Sinclair

[18] Tradução livre para o português: "Um corpus é definido em termos tanto de sua forma como de seu propósito. Lingüistas tem sempre usado a palavra corpus para descrever uma coleção de exemplos naturais da linguagem, consistindo de qualquer coisa desde algumas frases até um conjunto de textos escritos ou gravações de fitas que foram coletadas para um estudo lingüístico. Mais recentemente, a palavra tem sido reservada para coleções de textos (ou partes de texto) que são armazenadas e acessadas eletronicamente. Porque os computadores podem armazenar e processar grandes quantidades de informação, os corpora eletrônicos são usualmente maiores do que as coleções pequenas e baseadas em papel previamente utilizadas para estudar aspectos da linguagem. Um corpus é planejado apesar de o acaso poder desempenhar um papel na coleção de textos, e ele é projetado para algum propósito lingüístico. O propósito específico do planejamento determina a seleção de textos, e o objetivo é outro que não a preservação de textos propriamente ditos por causa de seus valores intrínsecos. Isto diferencia um corpus de uma biblioteca ou arquivo eletrônico. O corpus é armazenado de tal forma que pode ser estudado de forma não-linear, e tanto quantitativa como qualitativamente. O propósito não é simplesmente acessar os textos de forma a lê-los, o que novamente distingue o corpus de uma biblioteca ou arquivo".

[19] Tradução livre para o português: "um corpus é uma coleção de porções de textos em formato eletrônico, selecionados de acordo com critérios externos para representar, até onde for possível, uma língua ou uma varidade da língua como uma fonte de dados para a análise lingüística".

(2005), eletrônico, compilado com base em critérios externos, construído com o objetivo de pesquisa lingüística e podendo ser representativo de uma língua. Apesar de incluir a idéia de porções de textos na definição de corpus, o autor argumenta que "as amostras de linguagem para um corpus devem sempre que possível consistir de documentos inteiros ou transcrições de eventos de fala completos, ou devem chegar o mais próximo possível deste alvo"[20] (Sinclair, 2005).

Finalmente, tem-se a definição de Teubert & Čermáková (2007, p. 140): "*a collection of naturally occurring language texts in electronic form, often compiled according to specific design criteria and typically containing many millions of words*".[21] Esta, contudo, reitera elementos já apresentados em outras definições.

Assim sendo, nota-se a recorrência de alguns elementos definidores do corpus como indica a Figura 1, que os apresenta em ordem decrescente de menções.

[20] Tradução livre do seguinte fragmento: "Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible".

[21] Tradução livre para o português: "uma coleção de textos naturais em formato eletrônico, freqüentemente compilada de acordo com critérios específicos de planejamento e tipicamente contendo muitos milhões de palavras".

Obras / Termos chave	Tribble & Jones (1990)	Sinclair (1991)	Aarts (1991)	Murison-Bowie (1993)	McEnery & Wilson (1996)	Aston (1997)	Kennedy (1998)	Biber et al. (1999)	Tognini-Boneli (2001)	Conrad (2002)	Bowker & Pearson (2002)	Hunston (2002)	Sinclair (2005)	Teubert & Čermáková (2007)
Textos	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Compilação, coleção	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
Meio eletrônico				✓	✓	✓				✓	✓	✓	✓	✓
Seleção (criteriosa)		✓			✓				✓	✓	✓	✓	✓	✓
Naturalidade, autenticidade		✓							✓	✓	✓			✓
Representatividade	✓	✓			✓				✓				✓	
Amostra de textos			✓		✓							✓	✓	
Língua		✓							✓			✓	✓	
Propósito lingüístico							✓		✓			✓	✓	
Tamanho	✓				✓					✓	✓			
Gênero ou registro	✓							✓						
Autor	✓													
Documentos	✓													
Referência					✓									
Tópico	✓													

Figura 1: Elementos definidores do termo 'corpus' encontrados na literatura

Desta forma, tem-se que a associação de corpus com textos é a mais recorrente, não sendo mencionada em apenas uma definição. O corpus geralmente representa uma compilação de textos, aspecto mencionado em 12 das 14 definições analisadas. Recorrentemente faz-se menção à necessidade do corpus estar em formato eletrônico. Portanto, parece ser um pleonasmo utilizar a expressão 'corpus eletrônico' no estágio atual da lingüística de corpus visto que o meio eletrônico é constitutivo do reconhecimento de um corpus enquanto tal. Outras

noções também encontradas nas definições são a de seleção criteriosa, naturalidade dos textos e representatividade do corpus.

Nesta dissertação, algumas destas características são incluídas na seguinte definição do termo que expressa como o mesmo é aqui entendido: um corpus é uma compilação eletrônica e criteriosa de (amostras de) textos que ocorrem naturalmente com o objetivo de representar uma dada língua ou algum de seus aspectos mais pontuais de forma a possibilitar uma análise lingüística previamente delineada.

2.1.1. Tamanho

O tamanho de um corpus está intrinsicamente relacionado à representatividade do mesmo.[22] No entanto, não há um critério definido acerca do tamanho aceitável de um corpus para que ele seja representativo. Neste sentido, Bowker & Pearson (2002, p. 45) argumentam que a definição de corpus como uma grande coleção de textos não é clara o suficiente para especificar o que se entende pelo adjetivo 'grande'. Tognini-Bonelli (2001, p. 57), por exemplo, propõe a seguinte pergunta[23] a este respeito: "em relação ao tamanho, como podemos saber que um corpus não é grande o suficiente para os nossos propósitos?".[24] Isto, de certa forma, também se relaciona ao fato de que a questão da representatividade não é clara e explicitamente verificada. Em alguns casos o tamanho do corpus relaciona-se à disponibilidade de material que possa ser incluído no mesmo: "a questão do tamanho pode ser irrelevante, e isto não significa ser não-importante, no sentido de que o tamanho do corpus pode ser ditado pela quantidade de material que já está disponível em formato eletrônico ou

[22] A centralidade deste tópico na lingüística de corpus pode ser exemplificada com a existência de uma entrada para "tamanho" ("*size*") em um glossário dedicado a esta área (Baker, Hardie & McEnery, 2006).

[23] Perguntas semelhantes são encontradas em, por exemplo, Murison-Bowie (1993, p. 50) ("If big corpora are better than small ones, then how big is big?") e Baker, Hardie & McEnery (2006, p. 146) ("'how large should a corpus be?'"). De fato, a inclusão desta preocupação no manual do *MicroConcord* (Murison-Bowie, 1993) indica que esta controvérsia não é recente e nem foi ocasionada pela capacidade de armazenamento e processamento de dados cada vez maior por parte de computadores.

[24] Tradução livre do seguinte fragmento: "with regard to size, how can we ever know that a corpus is not large enough for our purposes?".

que tem que ser convertido ao mesmo"[25] (Pearson, 1998, p. 59). A mesma autora, entretanto, ressalta que o tamanho do corpus pode gerar problemas no caso de o mesmo ser tão pequeno que não seja considerado como representativo pela comunidade acadêmica.

Sinclair (2004, p. 189) adota uma postura mais firme na questão do tamanho de corpus, advocando o uso de grandes corpora na investigação da linguagem: "não há nenhuma virtude em ser pequeno. Pequeno não é belo; é simplesmente uma limitação."[26] O autor argumenta que mesmo que um corpus pequeno não seja reprovável – ou seja, ele tenha o potencial de fornecer os dados esperados – os resultados ainda assim são limitados, sendo melhor ter um corpus mais robusto que possa permitir a acomodação de maior variação no uso da língua (Sinclair, 2004).

Comumente aponta-se que o limite de tamanho para um corpus seria a capacidade de armazenamento e processamento da ferramenta computacional empregada para manter e analisar o mesmo (cf. Hunston, 2002). Assim sendo, de forma a tornar a análise factível são propostas duas soluções: (a) reduzir o tamanho do corpus, extraindo-se uma fração menor do mesmo para análise ou (b) trabalhar com o corpus por inteiro, mas selecionar de forma randômica as instâncias a serem analisadas. Parece ser a opção (b) a mais recorrente quando ambas as possibilidades estão disponíveis. De qualquer forma, deve-se ter em mente que "uma quantidade absoluta de informação pode ser assoberbante para o observador. [...] Poucos pesquisadores podem obter informação útil simplesmente olhando para tantas [143.000] linhas de concordância"[27] (Hunston, 2002, p. 25).

Quando da explicação acerca do planejamento da compilação de corpus, Biber, Conrad & Reppen (1998) indicam certas diretrizes com bases em pesquisas anteriormente realizadas pelo primeiro autor. Em relação ao número de textos a serem incluídos, deve-se ter, no mínimo, 10 textos visto que esta quantidade é

[25] Tradução livre do seguinte fragmento: "the issue of size may be irrelevant, and by that we do not mean unimportant, in the sense that corpus size may be dictated by the amount of material which is already available in electronic form or which has to be converted to electronic form".

[26] Tradução livre do seguinte fragmento: "There is no virtue in being small. Small is not beautiful; it is simply a limitation".

suficiente para representar muitas das categorias gramaticais. Em termos de número de palavras, indica-se que cada amostra de texto deva conter 1.000 palavras.

Apesar dos números reportados anteriormente, a explicação fornecida por Bowker & Pearson (2002, p. 45-46) parece ser apropriada no tocante a esta questão:

> Infelizmente, não há regras consistentes e seguras que possam ser seguidas para determinar o tamanho ideal de um corpus. Em vez disto, você terá que tomar esta decisão baseado em fatores como as necessidades do seu projeto, a disponibilidade de dados e a quantidade de tempo que você dispõe. É muito importante, no entanto, que não se suponha que maior é sempre melhor. Você pode achar que pode obter mais informações úteis de um corpus que é pequeno, mas bem planejado, do que de um que é maior, mas não é customizado para atender às suas necessidades.[28]

Em relação à classificação de tamanho de corpora, Berber Sardinha (2002) propõe uma escala baseada na análise de trabalhos completos publicados em cinco eventos dedicados à lingüística de corpus (três edições do ICAME [*International Computer Archive of Modern and Medieval English*], uma edição do PALC [*Practical Applications in Language and Computers*] e uma edição do TALC [*Teaching and Language Corpora*]).

Classificação do corpus	Número de palavras
Pequeno	Até 80 mil
Pequeno-médio	80 a 250 mil
Médio	250 mil a 1 milhão
Médio-grande	1 milhão a 10 milhões
Grande	Acima de 10 milhões

Figura 2: Proposta classificatória de corpora com base em tamanho (Berber Sardinha, 2002, p. 119)

[27] Tradução livre do seguinte fragmento: "Sheer quantity of information can be overwhelming for the observer. […] Few researchers can obtain useful information simply by looking at so many concordance lines".

[28] Tradução livre do seguinte fragmento: "Unfortunately, there are no hard and fast rules that can be followed to determine the ideal size of a corpus. Instead, you will have to make this decision based on factors such as the needs of your project, the availability of data and the amount of time that you have. It is very important, however, not to assume that bigger is always better. You may find that you can get more useful information from a corpus that is small but well designed than from one that is larger but is not customized to meet your needs".

A Figura 2 indica haver cinco possibilidades de classificação para corpora. No caso desta pesquisa, o corpus de estudo é descrito como pequeno (cf. Seção 5.1.1.5) ao passo que o de referência pode ser classificado como pequeno-médio (cf. Seção 5.1.2.6).

2.1.2. Tipologia

Há uma grande possibilidade classificatória de corpora dentro da lingüística de corpus visto que eles podem ser considerados por diferentes prismas. Dá-se preferência nesta seção para aqueles tipos que guardam alguma relação com o presente estudo.

Em termos de seu conteúdo, um corpus pode ser classificado como geral caso o objetivo seja o de representar uma língua como um todo, dando conta de suas diferentes nuances. Este é o caso, por exemplo, do BNC. O oposto corresponde a um corpus especializado que representa um aspecto mais pontual do uso de uma determinada língua. Isto é o que acontece com os corpora empregados nesta dissertação, que representam somente a redação produzida por universitários de Letras ou áreas afins, que já haviam cursado metade do tempo previsto para a integralização curricular.

Se o meio é considerado, um corpus pode ser escrito, contemplando textos deste tipo – como ocorre nesta dissertação; ou pode ser oral, contendo transcrições de fala. Ressalta-se que o meio escrito não necessariamente corresponde a textos impressos e/ou publicados, podendo contemplar textos manuscritos.

Caso a produção dos textos incluídos no corpus seja considerada a partir da questão temporal, o corpus pode ser sincrônico – representando um único período de tempo – ou diacrônico – representando diversos períodos de tempo. Se o tempo for o presente, pode-se descrever o mesmo como contemporâneo. Porém, se os textos tiverem sido produzidos em um passado mais distante, o corpus pode ser descrito como histórico. Os dois corpora deste estudo são considerados sincrônicos e contemporâneos.

Pode-se também considerar o conteúdo em termos da(s) língua(s) representada(s). Se um corpus contempla somente uma língua, como é o caso

aqui, ele é denominado de monolíngüe. No entanto, há corpora que contemplam mais de uma língua, sendo, por este motivo, denominados de multilíngües.

A questão da língua ainda pode ser refinada de forma a compreender quem são os produtores dos textos incluídos em um corpus. Se os textos são produzidos, por exemplo, em inglês por falantes de IL1 como no LOCNESS, este é descrito como sendo um corpus de língua nativa. No entanto, se o mesmo contempla textos em inglês escritos por falantes de ILE, este seria denominado de corpus de aprendiz (Granger, 1998a). Apesar de a utilização da palavra 'aprendiz', este termo é empregado para se referir a qualquer corpus com dados de falantes de LE não sendo obrigatório que os produtores sejam efetivamente alunos.

Não obstante a aceitação na literatura, talvez o termo 'corpus de aprendiz' precise ser revisto uma vez que ideologicamente ele colocaria todos os falantes de LE em uma posição de aprendizes eternos da qual eles nunca sairiam. Uma alternativa para esta questão talvez seja o emprego do termo 'corpus de L2 / LE', o que seria até mais congruente com o seu oposto – 'corpus de língua nativa'. No entanto, este, por sua vez, também poderia ser renomeado para 'corpus de L1' de forma a evitar o conceito de 'nativo'.

Por fim, a tipologia de um corpus pode estar relacionada ao seu objetivo. Caso planeje-se realizar a descrição de um corpus X em relação a um corpus Y, diz-se que o primeiro é de estudo enquanto o segundo é de referência. Uma vez que este é exatamente o objetivo da presente investigação, estes dois tipos de corpora podem ser identificados aqui.

2.2. Padrões de uso

Um ponto central da lingüística de corpus é a identificação de padrões no uso de uma língua. Esta área do conhecimento introduz a noção de que uma língua não se resume meramente ao preenchimento de espaços vazios de forma aleatória. Ressalta-se o papel da co-seleção exercido pelo cotexto,[29] ou seja, um determinado item apresenta preferência por outro item, padrão, etc em um

determinado ambiente lingüístico. No entanto, para se compreender a questão dos padrões de uso recorre-se aqui à primazia dos princípios advocados por Sinclair (1991). A Figura 3 ilustra, de forma resumida, como esta questão é entendida.

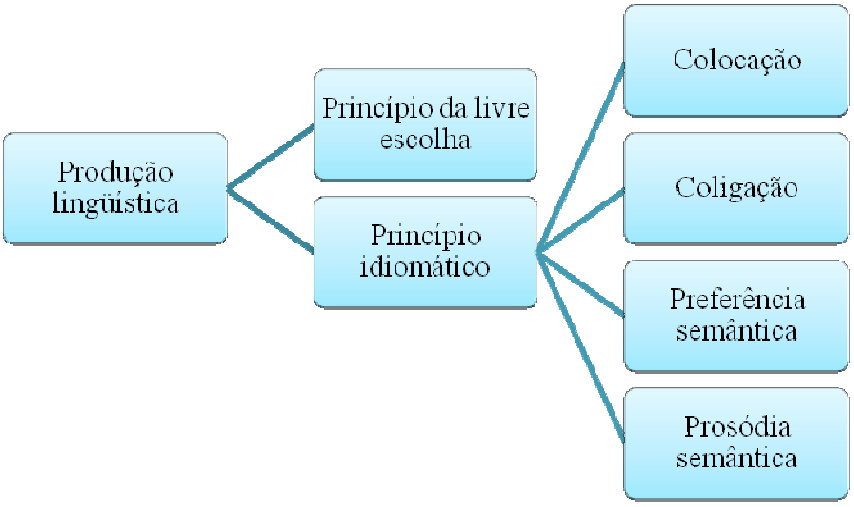

Figura 3: Produção lingüística à luz da lingüística de corpus (com base em Sinclair, 1991)

Toda a produção lingüística pode ser entendida a partir de dois princípios: o da livre escolha e o idiomático. De acordo com o primeiro, ao dizer ou escrever algo, seria possível escolher cada um dos itens, de forma individual, que compõem o enunciado. Nas palavras de Sinclair (1991, p. 109), "este é um modo de ver um texto como o resultado de um grande número de escolhas complexas".[30] O único aspecto a restringir esta liberdade seria a gramaticalidade, ou seja, as restrições impostas pela gramática de uma determinada língua.

O princípio idiomático, no entanto, é completamente diferente do primeiro uma vez que, neste caso, a produção lingüística seria guiada por padrões maiores do que simplesmente palavras. Assim sendo, "um usuário de uma língua tem à sua disposição um grande número de sintagmas semi-pré-construídos que constituem escolhas únicas, mesmo que eles pareçam ser analisáveis em

[29] Adota-se neste trabalho o termo 'cotexto' para se referir ao meio ambiente lingüístico. O termo 'contexto' é reservado para referências situacionais ou culturais conforme proposta da Gramática Sistêmico-Funcional (Halliday, 1994 [1985]).

[30] Tradução livre do seguinte fragmento: "This is a way of seeing language text as the result of a very large number of complex choices".

segmentos"[31] (Sinclair, 1991, p. 120). Em outras palavras, este princípio envolveria a seleção simultânea de estruturas maiores nas quais as palavras constituintes não estariam sujeitas à discriminação do falante e/ou escritor.

Pode-se argumentar que o princípio da livre escolha não introduz nenhum aspecto novo à investigação da linguagem visto que era esta a visão dominante antes da lingüística de corpus como inclusive aponta Sinclair (1991). A inovação de sua proposta reside justamente no princípio idiomático, possibilitando a abertura e a expansão de uma área de investigação denominada de fraseologia, que se ocupa da análise dos padrões de uso da linguagem.

A influência do princípio idiomático de Sinclair (1991) pode ser facilmente notada como colocam Viana, Silveira & Zyngier (2008). Por exemplo, o estudo de Erman & Warren (2002) verifica, de forma empírica, o impacto destes dois princípios na estrutura textual. Os resultados indicam haver uma grande ocorrência de expressões pré-fabricadas tanto em textos orais como escritos, o que "dá grande suporte ao princípio idiomático como formulado por Sinclair [...] e revelam que a proporção da pré-fabricação na linguagem tem geralmente sido muito subestimada"[32] (Erman & Warren, 2002, p. 50).

Nesta dissertação, interpreta-se o princípio idiomático como a pressuposição necessária para a compreensão dos padrões de uso de uma língua. Ressaltam-se aqui quatro destes padrões: colocações, coligações, preferências semânticas e prosódias semânticas conforme indicado na Figura 3.

A colocação, originalmente proposta por Firth (1957), pode ser identificada quando "as palavras parecem ser escolhidas em pares ou grupos e estes não são necessariamente adjacentes"[33] (Sinclair, 1991, p. 115). Assim sendo, este padrão de uso é eminentemente relacionado ao léxico, à co-escolha vocabular que é feita por falantes de uma língua. Em relação ao inglês, '*distinctly*' + '*unenthusiatic*' e '*hardly*' + '*surprising*' são dois exemplos de colocados (Hunston, 2002, p. 71).

[31] Tradução livre do seguinte fragmento: "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments".

[32] Tradução livre do seguinte fragmento: "give strong support to the idiom principle as formulated by Sinclair [...] and reveal that the proportion of prefabrication in language has generally been much underestimated".

[33] Tradução livre do seguinte fragmento: "words appear to be chosen in pairs or groups and these are not necessarily adjacent".

Há duas formas distintas de verificar os colocados de uma determinada palavra. Sinclair (2003) esclarece que alguns pesquisadores usam o termo somente para aquelas associações que são estatisticamente significativas, o que geralmente envolve testes como a informação mútua e os escores T ou Z (cf. McEnery & Wilson, 1996; Hunston, 2002; Berber Sardinha, 2004; McEnery, Xiao & Tono, 2006). Quando nota-se somente a ocorrência conjunta de palavras sem que a significância estatística seja verificada, opta-se pelo termo co-ocorrência (Sinclair, 2003).

Outra distinção entre tipos de colocados concerne àqueles denominados de 'colocados de coerência' ('*coherence collocates*') e os 'colocados vizinhos ('*neighborhood collocates*') (cf. Scott, 1998). Os primeiros relacionam-se às associações mentais. Por exemplo, 'hospital' provavelmente remeteria a 'médico', 'paciente' e/ou 'enfermeira'. Os colocados tidos como vizinhos são aqueles identificados na análise de corpus como recorrentes em um determinado horizonte que varia de quatro a cinco palavras em cada lado da palavra de busca (cf. Scott, 1998).[34] Sinclair (1991), no entanto, argumenta que este horizonte deve ser de quatro palavras à esquerda e quatro à direita.

Uma terceira distinção entre colocados é se o fenômeno é ascendente ou descendente (Sinclair, 1991). Todas as palavras de um dado corpus podem desempenhar duas funções: a de palavra nódulo e a de colocado. Assim sendo, se *x* é um colocado de *y*, pode-se considerar, em outro momento que *y* é colocado do nódulo *x*. Se considerarmos que *x* é mais freqüente do que *y*, o primeiro caso indicado (nódulo = *y*, colocado = *x*) é de colocação ascendente enquanto o inverso (nódulo = *x*, colocado = *y*) é um exemplo de colocação descendente. De acordo com Sinclair (1991), a verificação da colocação ascendente geralmente focaliza padrões gramaticais (por exemplo, '*at*', '*from*', '*into*', '*her*', '*his*', etc são colocados de '*back*') ao passo que a colocação descendente coloca padrões semânticos em evidência (como no caso de '*back*' ser colocado de '*arrive*', '*bring*', '*camp*', '*flat*', etc) (cf. Sinclair, 1991, p. 115-121).

[34] O autor adiciona que todos os dois tipos de colocados podem ser identificados com o auxílio do *WordSmith Tools* (Scott, 1999). Enquanto os colocados de coerência podem ser obtidos através da ferramenta *KeyWords*, os de vizinhança podem ser identificados com o *Concord* (cf. Seção 6.1.3).

Nesta dissertação, enfocam-se os colocados vizinhos descendentes, ou seja, aqueles que aparecem freqüentemente no cotexto de uma determinada palavra de busca, que é mais recorrente do que o colocado, identificados de forma manual a partir da análise das linhas de concordância (cf. Seção 6.1.3).

De forma semelhante à colocação, tem-se a coligação, que também se baseia em uma relação de co-ocorrência positiva, ou seja, ambos os elementos devem estar presentes. No entanto, enquanto a colocação opera em um nível lexical, a coligação ocorre em um nível gramatical ou léxico-gramatical. Esta é "a ocorrência de uma classe gramatical ou padrão estrutural com outro, ou com uma palavra ou sintagma"[35] (Sinclair, 2003, p. 173). A título de exemplificação, menciona-se que, em língua inglesa, '*flexible*' ocorre em posição atributiva com colocados abstratos ('*employment*', '*style*', '*ways*', '*options*', entre outros) – que são os mais freqüentes – e em posição predicativa nas poucas vezes que se refere a sujeitos animados (Tognini-Bonelli, 2001, p. 22).

Outro padrão de uso corresponde à preferência semântica. De acordo com Sinclair (2003, p. 178), "às vezes na estrutura de um sintagma há uma clara preferência por palavras de um significado particular. A classe gramatical não é importante e qualquer palavra com o significado apropriado será empregada".[36] O autor, no entanto, ressalta que há geralmente a ocorrência de padrões colocacionais observáveis também dentro da preferência semântica, o que faz com que algumas palavras sejam mais recorrentemente empregadas em presença de uma determinada palavra do que outras que são classificadas dentro da mesma preferência semântica. Um exemplo de preferência semântica ocorre no uso de '*completely*' com alguma palavra indicativa de ausência ('*disappeared*', '*empty*', '*hopeless*', '*gone*', etc) ou de mudança ('*altered*', '*changed*', '*destroyed*' e '*different*') (cf. Partington, 2004).

Por fim, há a possibilidade de se identificar a prosódia semântica, um conceito que se aproxima da pragmática (Tognini-Bonelli, 2001). Segundo Sinclair (2003, p. 178), a prosódia se refere a "um evento significativo que não é necessariamente localizado em uma unidade de expressão particular, mas que se

[35] Tradução livre do seguinte fragmento: "the occurrence of a grammatical class or structural pattern with another one, or with a word or phrase".

[36] Tradução livre do seguinte fragmento: "Sometimes in the structure of a phrase there is a clear preference for words of a particular meaning. The word class is not important, and any word with the appropriate meaning will do".

estende por várias".[37] É exatamente esta noção de extensão que difere a prosódia semântica da conotação: enquanto a primeira é obrigatoriamente colocacional, a segunda pode ser ou não desta natureza (McEnery, Xiao & Tono, 2006, p. 85).[38]

A prosódia semântica é definida como "uma aura consistente de significado com a qual uma forma é imbuída por seus colocados"[39] (Louw, 1993, p. 157). Este padrão assume, na maioria das vezes, um valor negativo, sendo que somente em alguns casos há a ocorrência de prosódias semânticas positivas (McEnery, Xiao & Tono, 2006). Exemplifica-se este fenômeno com o uso da palavra '*utterly*', que implica uma prosódia semântica negativa visto que é empregada com '*blackened*', '*burned*', '*confused*', '*exaustive*', entre outras. Louw (1993) indica que só há quatro exceções a este padrão no corpus de 18 milhões de palavras do Cobuild no qual '*utterly*' estaria aparentemente empregado com colocados indicativos de positividade como '*dedicated*', '*good*', '*grand*' e '*venerable*'. Porém, todas elas são empregadas em um contexto irônico, o que vem a confirmar a prosódia semântica negativa de '*utterly*'. A violação de uma prosódia semântica seria realizada, segundo Louw (1993), para gerar algum tipo de efeito no leitor / escritor seja ele de ironia, insinceridade ou humor.

Deve-se distinguir claramente a 'preferência semântica' da 'prosódia semântica'. No primeiro caso, o que está em jogo é o próprio significado dos colocados, sendo a relação entre palavras o aspecto mais importante. No tocante à prosódia semântica, enfoca-se o significado afetivo de uma palavra com os seus colocados. Apesar de os conceitos serem diferentes, eles são interdependentes (McEnery, Xiao & Tono, 2006). Para Sinclair (2003), é a revelação de prosódias semânticas que se constitui em uma das maiores contribuições da lingüística de corpus ao estudo da linguagem.

[37] Tradução livre do seguinte fragmento: "a meaningful event that is not necessarily located in a particular unit of expression, but may spread over several".

[38] Por vezes, considera-se a conotação como uma característica intrínseca da palavra.

[39] Tradução livre do seguinte fragmento: "A consistent aura of meaning with which a form is imbued by its collocates".

2.3. Resumo

Este capítulo enfocou os aspectos mais importantes da lingüística de corpus tendo em vista o estudo aqui realizado. Foram comentadas as especificidades dos corpora assim como suas possíveis classificações.

Ressaltou-se que a lingüística de corpus e exige uma análise calcada em dados empíricos que permitam o acesso ao emprego que é feito de uma língua por seus falantes. É exatamente a partir da observação de muitos textos com o auxílio computacional que se pode notar mais facilmente a ocorrência de colocações, coligações, preferências semânticas e prosódias semânticas, revelando o aspecto fraseológico do uso lingüístico.

Como Sinclair (1997) aponta na epígrafe deste capítulo, é preciso concentrar-se em exemplos reais de uso seja no ensino – como ocorre no contexto do qual a citação foi retirada – ou na pesquisa – como é o caso aqui. Não é possível inventar fatos lingüísticos, pode-se somente observá-los. Em outras palavras, talvez a máxima da lingüística de corpus possa ser resumida com o título de uma coletânea de textos de Sinclair (2004): confie no texto.