D.E.L.T.A., Vol. 16, Nº. 2, 2000 (323-367)

RETROSPECTIVA/RETROSPECTIVE

LINGÜÍSTICA DE CORPUS: HISTÓRICO E PROBLEMÁTICA
(Corpus Linguistics: History and Problematization)

Tony Berber SARDINHA
(LAEL, PUC-SP)

ABSTRACT: This paper offers an overview of Corpus Linguistics, which is a research area that has experienced a considerable growth in the past years and which has made a considerable impact on linguistics. The overview looks at both the past and the present of Corpus Linguistics. The main concepts in the area are presented and debated, and the paper also comments on the main theoretical aspects in the field. The principal corpora and software are reviewed.
KEY-WORDS: Corpus Linguistics, corpora, history of Corpus Linguistics, theory of Corpus Linguistics.

RESUMO: O presente trabalho oferece uma retrospectiva da Lingüística de Corpus, uma área de pesquisa que tem experimentado um crescimento vertiginoso nos últimos anos e que tem tido um impacto considerável na lingüística. A retrospectiva inclui tanto um painel histórico quanto um posicionamento em relação aos debates correntes e desenvolvimentos futuros da área. Os conceitos principais em voga na área são apresentados e discutidos. O trabalho ainda comenta os fatos mais marcantes na Lingüística de Corpus em relação à teoria e à prática, elencando os principais corpora *em existência bem como as mais importantes contribuições no campo de programas de computador para análise e exploração desses corpora.*
PALAVRAS-CHAVE: Lingüística de Corpus, corpora, *história da Lingüística de Corpus, teoria da Lingüística de Corpus.*

1. Introdução

No ano de 1999 comemorou-se o aniversário de 35 anos da criação do primeiro corpus lingüístico eletrônico, o corpus Brown. Lançado em 1964, o Brown University Standard Corpus of Present-Day American

English, continha uma quantidade invejável de dados para a época: um milhão de palavras. Há 35 anos as dificuldades de se informatizar um conjunto de textos eram tremendas. Vale lembrar, por exemplo, que os textos tiveram de ser transferidos para o computador por meio de cartões, perfurados um a um, tal era a tecnologia da época. Este feito, por si só, já traria respeito e admiração à empreitada.

Mas não foi somente o pioneirismo[1] que garante uma posição de destaque para o corpus Brown. Há também a conjuntura histórica. O corpus Brown foi lançado justamente numa época em que a idéia de se gastar tempo e recursos financeiros para a coleta de registros lingüísticos era vista com total incredulidade e até hostilidade. Lembremo-nos de que há apenas 7 anos havia sido lançado 'Syntactic Structures', obra de Noam Chomsky, que teria papel fundamental em nada menos do que uma mudança de paradigma na lingüística. Dentro desta visão de linguagem, que se instauraria a partir desta obra de Chomsky, os dados necessários para o lingüista estavam em sua mente e eram acessíveis por meio da introspecção. Não havia necessidade de coletar-se dados abundantes de terceiros. Estes serviriam apenas para o estudo do desempenho, quando todos sabiam que o que interessava era a investigação da competência lingüística. Portanto, o corpus Brown surgira numa época em que seu mérito era discutido.

Esta nota histórica tem a função não só de homenagear o corpus Brown como tal, mas também (e principalmente) de salientar sua importância enquanto fato que impulsionou o desenvolvimento da área conhecida atualmente por Lingüística de Corpus, uma das áreas de pesquisa de linguagem mais ativas nos últimos anos[2]. Não que ela não existisse não fosse o corpus Brown, mas com certeza seria muito diferente. Este artigo irá se ocupar não desse corpus , em particular, mas da Lingüística de Corpus em geral. O objetivo do trabalho é aproveitar o ensejo da comemoração do aniversário do corpus Brown para fazer uma

[1] Entendido aqui em relação a corpora de linguagem escrita. O primeiro corpus eletrônico de linguagem falada, com 220 mil palavras, é atribuído a John McH. Sinclair (vide Sinclair, 1995, p. 99).

[2] A discussão acerca de se a Lingüística de Corpus é uma disciplina ou metodologia será apresentada na seção 6.

retrospectiva da Lingüística de Corpus, na qual se pretende apresentar os principais marcos na sua história, como também discutir algumas questões teóricas e práticas subjacentes a ela. A maior parte do texto será dedicada ao processamento da língua inglesa, visto que é em relação a esta língua que se deu o maior desenvolvimento na área.

2. A Lingüística de Corpus e seu histórico

A Lingüística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.

Havia corpora antes do computador, já que o sentido original da palavra 'corpus' é 'corpo', 'conjunto de documentos' (conforme o dicionário Aurélio). Na Grécia Antiga, Alexandre, o Grande definiu o Corpus Helenístico. Na Antiguidade e na Idade Média, produziam-se corpora de citações da Bíblia.

Durante boa parte do século XX houve muitos pesquisadores que se dedicaram à descrição da linguagem por meio de corpora, entre eles educadores como Thorndike e lingüistas de campo como Boas. Há duas diferenças fundamentais entre esta época e a atual. A primeira, obviamente, é que os corpora não eram eletrônicos, ou seja, eram coletados, mantidos e analisados manualmente. A segunda é que a ênfase destes trabalhos era em geral o ensino de línguas. Atualmente o que prepondera na literatura é a descrição de linguagem e não a pedagogia, embora recentemente tenha ressurgido um interesse no emprego de corpora na sala-de-aula e na investigação da linguagem de alunos de língua (Granger, 1998).

Um trabalho fenomenal, dada as condições da época, foi a identificação das palavras mais freqüentes da língua inglesa, feita por Thorndike há mais de 75 anos atrás (Thorndike, 1921). O levantamento foi feito manualmente em um corpus de nada menos de 4,5 milhões de palavras,

e, quando publicado, impulsionou mudanças no ensino de língua materna e estrangeira, tanto nos Estados Unidos quanto na Europa. As abordagens baseadas no controle do vocabulário, nas quais os alunos têm contato em primeiro lugar com as palavras mais freqüentes,devem sua inspiração a estudos como o de Thorndike. Quase 25 anos mais tarde, Thorndike revisou seu levantamento inicial e, tomando como base um corpus maior, com impressionantes 18 milhões de palavras, publicou uma obra listando as 30 mil palavras mais comuns da língua inglesa. Logo depois, em 1953, veio o 'General Service List of English Words' de Michael West (West, 1953), talvez a mais famosa descrição do léxico inglês pré-computador. A pesquisa de West dá detalhes do que seriam as 2 mil palavras mais freqüentes do inglês e baseou-se no trabalho de pioneiros como Thorndike e Lorge.

Foi um corpus não computadorizado que deu feição aos corpora atuais, o SEU (Survey of English Usage), compilado por Randolf Quirk e sua equipe, em Londres, a partir de 1953. O SEU foi planejado para ter o tamanho de 1 milhão de palavras, depois tido como referência por outros corpora, inclusive o Brown. A composição do corpus também foi influente, ao definir um número fixo de textos (200) e uma quantidade de palavras igual para cada texto (5000). O *Survey* foi organizado em fichas de papel, cada um contendo uma palavra do corpus inserida em 17 linhas de texto. As palavras foram analisadas gramaticalmente, com cada ficha recebendo uma categoria gramatical. O conjunto de categorias resultante serviu de base para o desenvolvimento dos etiquetadores computadorizados contemporâneos, que fazem a identificação de traços gramaticais automaticamente. A famosa *Comprehensive Grammar of the English Language* de Quirk, Greenbaum, Leech e Svartvik foi baseada no SEU. A transformação completa do *Survey* em corpus eletrônico só foi atingida muitos anos depois, em 1989, mas a sua parte falada foi computadorizada antes e ficou conhecida como o London-Lund Corpus.

No final dos anos 50 apareceria 'Syntactic Structures', de Chomsky, e com ele uma mudança de paradigma na lingüística: saía de cena o empirismo e com ele a sustentação dos trabalhos baseados em corpora, tomando lugar central as teorias racionalistas da linguagem (vide dis-

cussão abaixo), notadamente a lingüística gerativa. Além do apelo natural da lingüística Chomskyana, outro fator que contribuiu para a perda de fôlego de abordagens baseadas em corpus foi uma crescente leva de críticas sobre o processamento manual de corpora. Uma das críticas mais contundentes era exatamente que o processamento de corpora gigantescos, como o de Thorndike, com 18 milhões de palavras, por meios manuais, não era confiável. O ser humano não é talhado para tarefas deste tipo. Não seria o caso de simplesmente aumentar a equipe de analistas para resolver o problema, pois este trabalho já era realizado com grandes contingentes de assistentes. A pesquisa de Käding, por exemplo, sobre a ortografia do alemão, consumiu a mão-de-obra de 5000 analistas! Os problema da possibilidade de erro e de falta de consistência persistem, ou até pioram, com grandes equipes. A outra alternativa era diminuir o tamanho dos corpora para facilitar a inspeção manual, mas isto atentava contra a própria natureza da pesquisa. O que faltava era justamente um instrumento que permitisse a análise de grandes quantidades de dados de modo confiável, mas a tecnologia da época não permitia isso.

A invenção do computador mudou este quadro. Nos anos 60, os computadores *mainframe* passaram a equipar centros de pesquisa universitários e foram sendo aproveitados para a pesquisa em linguagem. Com a popularização dos computadores, foi possibilitado o acesso de mais pesquisadores ao processamento de linguagem natural e, concomitantemente, a sofisticação do equipamento permitiu a consecução de tarefas mais complexas, mais eficientemente, sem falar no aumento da capacidade de armazenamento e na introdução de novas mídias (fitas magnéticas, em vez de cartões *hollerith* perfurados, etc.), as quais facilitaram a criação e manutenção de corpora em maior número. Com a entrada em cena dos micro-computadores pessoais, nos anos 80, uma nova onda de mudanças aconteceu, como a popularização de corpora e de ferramentas de processamento, o que contribuiu decisivamente para o reaparecimento e fortalecimento da pesquisa lingüística baseada em corpus.

Hoje em dia, a Lingüística de Corpus é de grande influência na pesquisa lingüística, em vários centros. Na Grã-Bretanha, um dos cen-

tros mais desenvolvidos, várias universidades (Birmingham, Brighton, Lancaster, Liverpool, Londres, etc.) dedicam-se à pesquisa baseada em corpus para a descrição dos mais variados aspectos da linguagem. A pesquisa em instituições britânicas tem possibilitado tanto a teorização quanto a criação de corpora e de materiais de apoio em diversas áreas. Igualmente, nos países escandinavos (Noruega, Suécia e Dinamarca) existem centros estabelecidos dedicados à Lingüística de Corpus com um papel atuante há vários anos.

Fora da Europa, a Lingüística de Corpus não está tão desenvolvida, mas já possui centros nos quais a pesquisa está instalada. Paradoxalmente, nos Estados Unidos, tendo-se em vista a pujança de seus centros de pesquisa e a facilidade de obtenção de recursos de informática, a Lingüística de Corpus tem uma presença mais modesta. Uma explicação é a força da lingüística gerativa-transformacional nos departamentos de lingüística, a qual conflita naturalmente com a Lingüística de Corpus. Evidência disto é que um dos maiores expoentes da Lingüística de Corpus mundial, o americano Douglas Biber, atua em um departamento de inglês. Por outro lado, há nos Estados Unidos um alto estágio de desenvolvimento na pesquisa em Processamento de Linguagem Natural (PLN), tanto em nível acadêmico quanto industrial (as empresas de informática investem pesado na pesquisa lingüística com fins comerciais). O Processamento de Linguagem Natural é uma disciplina com laços fortes com a Ciência da Computação e, embora compartilhe vários temas com a Lingüística de Corpus, as duas mantém-se independentes.

No Brasil, a Lingüística de Corpus ainda é incipiente. A pesquisa em corpus se dá em centros mais voltados ao Processamento de Linguagem Natural, Lexicografia e à lingüística Computacional (vide Berber Sardinha, 1999).

Não é só nos centros acadêmicos que a Lingüística de Corpus tem ganhado espaço. Também no âmbito empresarial tem havido um interesse crescente nas aplicações comerciais de estudos baseados em corpora. Primeiramente, deve-se destacar as parcerias entre empresas e universidades. Aqui a norma é a associação de um centro de pesquisa

em Lingüística de Corpus com uma editora. O pioneiro neste sentido é o COBUILD, uma parceria entre a Universidade de Birmingham (Grã-Bretanha) e a editora Collins. No âmbito do COBUILD foram produzidos vários dicionários, gramáticas e livros didáticos para o ensino do inglês. Atualmente quase desativado, o COBUILD permanece como referência no desenvolvimento e aplicação da pesquisa baseada em corpus com fins comerciais.

Os principais membros do COBUILD vieram a fundar ou a se incorporar a outros centros. Antoinette Renouf, por exemplo, pesquisadora sênior no projeto COBUILD, veio a instituir a Unidade de Pesquisa e Desenvolvimento junto à Universidade de Liverpool (Grã-Bretanha), que se dedica a parcerias entre as empresas e a universidade. Parcerias semelhantes ao COBUILD entre empresas e universidades britânicas hoje são comuns, notadamente voltadas para a produção de dicionários, como por exemplo entre o grupo Addison-Wesley/Longman e a universidade de Lancaster (Grã-Bretanha).

Em segundo lugar, há um desenvolvimento crescente de centros de pesquisa mantidos por empresas. Estes centros utilizam-se de pesquisas baseadas em corpus para várias finalidades comerciais, como o processamento automático de textos, informatização de grandes bases de dados e a montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento de informação. As grandes empresas de telecomunicações investem nestas áreas, reconhecendo o potencial econômico deste campo. Outras empresas de produtos de informática como a Xerox, Microsoft e Canon também possuem centros desenvolvidos de pesquisa de corpus e Processamento de Linguagem Natural.

A história da Lingüística de Corpus está, portanto, intimamente ligada à disponibilidade de corpora eletrônicos. Os principais corpora compilados, ou em compilação, até hoje são:

Corpus	Lançamento / Referência na literatura	Palavras	Composição
Brown Corpus (Brown University Standard Corpus of Present-Day American English)	1964	1 milhão	Inglês americano, escrito
AHI (American Heritage Intermediate Corpus)	1971	5 milhões	Inglês americano, escrito
LOB (Lancaster-Oslo-Bergen)	1978	1 milhão	Inglês britânico, escrito
LLC (London-Lund Corpus)	1980	500 mil	Inglês britânico, falado
Birmingham Corpus (Birmingham University International Language Database)	1987	20 milhões	Inglês britânico
Kolhapur Corpus (of Indian English)	1988	1 milhão	Inglês indiano, escrito
TOSCA Corpus (Tools for Syntactic Corpus Analysis)	1988	1,5 milhão	Inglês britânico, escrito
SEU Corpus (Survey of English Usage)	1989	1 milhão	Inglês britânico, escrito e falado
CHILDES (Child Language Data Exchange)	1990	20 milhões	Inglês infantil, falado
Nijmengen Corpus	1991	132 mil	Inglês britânico, escrito e falado
LLELC (Longman-Lancaster English Language Corpus)	1991*	50 milhões*	Inglês de vários tipos, escrito e falado
Map Task Corpus	1991	147 mil	Inglês escocês, falado
LCLE (Longman Corpus of Learner's English)	1992	10 milhões	Inglês escrito por estrangeiros
SEC (Lancaster/IBM Spoken English Corpus)	1992	53 mil	Inglês britânico, falado
Wellington Corpus (of Written New Zealand English)	1993	1 milhão	Inglês neozelandês, escrito
POW (Polytechnic of Wales Corpus)	1993	65 mil	Inglês infantil, falado
Wellington Corpus of Spoken New Zealand English	1995	1 milhão	Inglês neozelandês, falado
BNC (British National Corpus)	1995	100 milhões	Inglês britânico, escrito e falado
Corpus of Spoken American English	1991	2 milhões	Inglês americano, falado
ICLE (International Corpus of Learner English)	1997	200 mil**	Inglês escrito por estrangeiros
Bank of English	1997	320 milhões	Inglês britânico

* previsão

** cada variedade nacional

A partir da tabela anterior, pode-se perceber três corpora eletrônicos que servem como marcos de referência históricos: Brown, Birmingham e BNC. O corpus Brown é um marco por razões óbvias: é o pioneiro. O corpus Birmingham é importante porque foi o primeiro a ultrapassar a marca de 1 milhão de palavras iniciada pelo Brown. Vale lembrar que o corpus Birmingham se tornaria o Bank of English, sempre em crescimento, atingindo agora 320 milhões de palavras. Por fim, o BNC é um marco histórico porque foi o primeiro a conter 100 milhões de palavras e ainda é, dentre os mega-corpora, o único disponível para compra (dentro da Comunidade Européia apenas). O Bank of English é de acesso restrito aos pesquisadores ligados ao COBUILD e à editora Collins.

Os corpora elencados acima são de língua inglesa, mas há corpora de várias outras línguas: francês (http:// hydre.auteuil. cnrs-dir.fr/ cnrseditions/ sources/ sTlf.asp), espanhol (Sanchez et al., 1995), alemão (http:// corpora.ids– mannheim.de/ ~cosmas), tcheco (Cermak, 1997), chinês (Zhou e Yu, 1997) e Estoniano (Hennoste et al., 1998), para mencionar apenas algumas.

Na língua portuguesa, há vários corpora eletrônicos de destaque, tais como o Corpus de Araraquara, o de São Carlos (NILC), o CRPC (Corpus de referência do português contemporâneo), o Banco de Português, o PORTEXT, o Tycho-Brahe (português histórico) e o Corpus Natura, para citar apenas alguns. A pesquisa com corpora eletrônicos no Brasil já vem de longa data. Biderman (1978, pp.265-266) cita o corpus do 'Frequency Dictionary of Portuguese Words' como um dos primeiros corpora eletrônicos de português. Esse corpus continha 500 mil palavras de português europeu, referentes a publicações de 1920 a 1940. O dicionário de freqüências feito a partir dele foi concluído em 1972 mas permanece inédito (Duncan Jr, 1972). Biderman (1978, pp. 65-67) ainda menciona vários outros corpora pioneiros no Brasil, usados para pesquisas no campo da Estatística Léxica, destacando-se os compilados por Jean Roche (Universidade de Toulouse, França, na década de 1960), J. Hutchins (Academia Naval de Anápolis, EUA, anos 1970), Cléa Rameh (Universidade Stanford, EUA, 1972), além daquele compilado por ela mesma (Maria Teresa Biderman, USP, 1969) e de uma série de corpora de textos literários de autores brasileiros construídos

e analisados por uma equipe do ITA (São José dos Campos). Castilho et al. (1995) oferecem um panorama dos projetos de criação e informatização de corpora em várias regiões do Brasil. O levantamento indicou que havia um interesse na criação de corpora por parte de vários grupos de pesquisa, embora o índice de informatização estivesse apenas pouco acima de 50%. Castilho et al. (1995) concluem pormenorizando o que seria o Banco de Dados da Língua Portuguesa, um corpus de língua escrita e falada, o qual não foi concretizado.

Esses e outros corpora proporcionaram o acúmulo de uma extensa obra em Lingüística de Corpus, cujos principais marcos, a nosso ver, são os seguintes:

- Sinclair, 1966. O trabalho pioneiro na área de léxico que traçou os caminhos da maioria da pesquisa em Lingüística de Corpus feita até hoje.
- Leech, 1966. O primeiro trabalho sobre análise de corpus publicado por Geoffrey Leech, um dos maiores lingüistas de corpus de todos os tempos, no qual ele antecipa a necessidade de análises detalhadas de corpora via computador.
- Francis, W. N. e Kucera, 1982. A descrição por computador das freqüências do pioneiro dos corpora, o Brown.
- Sinclair et al., 1987. Lançamento do dicionário COBUILD, o primeiro a ser compilado a partir de um corpus computadorizado. Seus verbetes e definições foram compostos com informações provenientes do corpus. Hoje em dia, o emprego de corpora na produção de dicionários, em língua inglesa pelo menos, tornou-se rotineira.
- Aijmer e Altenberg, 1991. A primeira grande obra que adota a expressão 'Corpus Linguistics' no título.
- Svartvik, 1992. A academia de ciências da Suécia dedica um de seus célebres seminários 'Nobel' à Lingüística de Corpus. Os mais renomados lingüistas da época comparecem para apresentar um painel do estado da arte naquele momento.
- Biber, 1988. O trabalho monumental de descrição da composição lingüística de gêneros da língua inglesa a partir de dois dos mais famosos corpora (LOB e London-Lund) abriu os olhos de muitos pesquisadores para a necessidade da investigação do texto. O autor, hoje um

dos mais atuantes na Lingüística de Corpus, não se intitulava 'lingüista do corpus' então.

- Sinclair, 1991. O maior lingüista de corpus da história reúne alguns de seus trabalhos principais em uma obra que encerra muitas das idéias centrais da área em aplicação até hoje, notadamente 'colocação'.
- Kjellmer, 1994. Primeiro dicionário de colocações baseado em corpus (no caso, o próprio Brown), elaborado a partir de padrões recorrentes identificados estatisticamente. O seu predecessor, o dicionário BBI de colocações (Benson et al., 1986), não dá indicação clara de ter sido criado seguindo os mesmo princípios.
- McEnery e Wilson, 1996. Um manual de Lingüística de Corpus de tom didático e com ampla cobertura de conceitos práticos e teóricos. Ao contrário dos trabalhos anteriores, os quais se voltavam a pesquisadores formados, dedica-se a alunos de Lingüística de Corpus, um dos nichos mais importantes da área. Denota a expansão da área.
- Francis, G. e Hunston, 1996. Primeira 'gramática do léxico', descreve de modo amplo e profundo os padrões verbais da língua inglesa a partir de um corpus, seguindo o princípio básico da identificação de colocações recorrentes por computador. O segundo volume foi lançado a seguir, dedicado aos substantivos e adjetivos (Francis, G. e Hunston, 1998). A formulação teórica dos princípios seguidos nas gramáticas apareceu mais recentemente em Hunston e Francis (2000).
- Biber et al., 1998. Este outro manual de Lingüística de Corpus proporciona uma perspectiva americana da área que até então era dominada exclusivamente por trabalhos provenientes de centros de pesquisa europeus.
- Granger, 1998. Coletânea que reúne trabalhos voltados a uma das áreas que mais crescem: a aplicação de corpus no ensino e na aprendizagem de línguas. Também consolida um tipo de corpus diferente dos demais, o corpus de aprendizes, formado por amostras de falantes não-nativos.
- Partington, 1998. Volta-se diretamente ao praticante da Lingüística de Corpus 'caseira', isto é, aqueles que trabalham com computadores pessoais e corpora pequenos. Fala mais diretamente ao professor de línguas e ao tradutor.

Além dessas obras específicas, a cronologia de outros veículos importantes de divulgação da pesquisa da área é:

- 1979. Primeira conferência ICAME, até hoje o fórum regular mais importante da área. Ainda é um evento exclusivo, onde os participantes são convidados.
- 1994. Primeira conferência bienal TALC (Teaching and Learning Corpora), especializada na aplicação de corpora no ensino e aprendizagem de línguas.
- 1997. Primeira conferência PALC (Practical Applications of Language Corpora). Inspirada no sucesso da TALC, leva a Lingüística de Corpus para fora da Europa ocidental, favorecendo os pesquisadores do antigo bloco comunista que há muito se dedicavam a questões de lingüística de corpus.
- 1996. Primeira edição do International Journal of Corpus Linguistics, o primeiro (e até agora único) periódico dedicado exclusivamente à Lingüística de Corpus.
- 1998. Lançamento do primeiro volume da série 'Studies in Corpus Linguistics' da editora Benjamins, a primeira série de livros que se faz valer do rótulo 'Lingüística de Corpus'.

A história da Lingüística de Corpus está condicionada à tecnologia, que permite não somente o armazenamento de corpora, mas também a sua exploração. Por isso, a história da área está relacionada à disponibilidade de ferramentas computacionais para análise de corpus, dentre as quais se destacam as seguintes:

- 1970. TAGGIT, o primeiro etiquetador morfossintático para computador.
- 1979. CLAWS, o etiquetador mais famoso em utilização, usado na sua forma atual para anotar o BNC (British National Corpus). Roda em mainframes.
- 1987. TACT. Um dos programas pioneiros para micro-computadores, permite a consecução das tarefas principais de análise de corpus (listagem de palavras e concordâncias).

- 1988. OCP. The Oxford Concordance Program, um dos principais concordanceadores usados em microcomputadores e estações de trabalho.
- 1993. MicroConcord. O mais famoso, simples e robusto programa de concordância para micro-computadores, até hoje.
- 1995. WordSmith Tools. Primeiro a aproveitar os recursos do ambiente Windows para análise de corpus, divulga a Lingüística de Corpus entre usuários de micro-computadores. Ainda hoje, depois de muitas versões, o mais completo e versátil conjunto de ferramentas para Lingüística de Corpus.
- 1997. Brills tagger para DOS. Versão para micro-computadores do etiquetador Brill, um dos mais famosos e mais facilmente disponíveis para a comunidade. Permite que o usuário de micro-computadores faça, além de contagens de palavras e concordâncias, a etiquetagem de seu corpus. Para a maioria dos usuários sem grande conhecimento técnico, restringe-se a corpora em inglês.
- 1998. QTAG. A etiquetagem entra na era multiplataforma com este etiquetador para Java. Agora o mesmo etiquetador para grandes máquinas roda em micros também. Além disso, quebra a hegemonia da etiquetagem do inglês, pois permite que o usuário treine o etiquetador para análise de outras línguas.

3. Corpus: Tipologia, Representatividade e Extensão

Central à Lingüística de Corpus atual é a existência de uma coletânea de dados lingüísticos naturais, legíveis por computador. Mas nem todo conjunto de dados é considerado um corpus:

- Arquivo: depósito de textos sem organização prévia;
- Biblioteca eletrônica: Coleção que segue alguns critérios de seleção;
- Corpus: Uma parte da biblioteca eletrônica, construído a partir de um desenho explícito, com objetivos específicos;
- Sub-corpus: Uma parte de um corpus, pode ser fixa ou mutável (dinâmica, i.e. flexível durante a análise) (Atkins et al., 1992, p. 1)

Proliferam na literatura definições de corpus. Algumas das mais importantes são apresentadas a seguir.

> 'Uma coletânea de textos naturais ('naturally occurring'), escolhidos para caracterizar um estado ou variedade de linguagem'. (Sinclair, 1991, p. 171).

Por textos naturais entende-se 'autênticos', isto é, aqueles que existem na linguagem e que não foram criados com o propósito de figurarem no corpus. Além disso, amplia-se a idéia de 'natural' para incluir somente aqueles textos produzidos por humanos. Desta forma está excluída a produção provinda de programas de geração de textos.

Um problema com esta definição é que ela não deixa claro o propósito da criação do corpus. Por isso, deve-se incorporar a ela a complementação abaixo:

> '[Corpus é] um corpo de linguagem natural (autêntica) que pode ser usado como base para pesquisa lingüística'. (Sinclair, 1991, p. 171)

Agora estabelece-se que um corpus é um artefato produzido para a pesquisa. Assim, se por um lado os textos devam ser naturais (autênticos e independentes do corpus), o corpus em si é artificial, um objeto criado com fins específicos de pesquisa. Estes dois posicionamentos estão presentes na conceituação abaixo:

> 'Corpus é uma coletânea de porções de linguagem que são selecionadas e organizadas de acordo com critérios lingüísticos explícitos, a fim de serem usadas como uma amostra da linguagem'. (Percy et al., 1996, p. 4).

É importante destacar na definição o termo 'porções de linguagem', empregado em lugar de 'textos'. Isto se deve ao fato dos problemas relacionados à delimitação do conceito de 'texto', já que se pode considerar tanto um artigo científico, quanto o seu resumo inicial, quanto um trecho de conversação, como texto. Por isso se fala aqui em porções de linguagem, um conceito que acomoda estas três instâncias.

Por não seguirem estes preceitos, a definição a seguir é inadequada:

> 'Um corpo de material lingüístico que existe em formato eletrônico e que pode ser processado por computador para vários propósitos.' (Leech, 1997, p. 1)

Esta definição permitiria que qualquer conjunto de textos eletrônico fosse considerado um corpus. Mas conforme dito antes, um corpus deve ser planejado e concretizado seguindo critérios lingüísticos de seleção.

A definição a seguir também é inapropriada:

> 'Corpus de material lingüístico natural (textos inteiros, amostra de textos, ou às vezes somente sentenças desconexas), que são armazenadas em formato legível por máquina'. (Leech, 1991, pp. 115-116)

Esta definição permite não somente que qualquer coletânea eletrônica seja um corpus, mas que também qualquer conteúdo eletrônico o seja, tais como sentenças soltas. A princípio, a linguagem natural autêntica não é formada de fragmentos desconexos e, portanto, sentenças soltas não seriam representantes da linguagem. A exceção seria se o corpus fosse criado exatamente para ser uma coletânea de frases soltas.

A definição a seguir faz menção à extensão do corpus:

> 'Uma coletânea grande e criteriosa de textos naturais' (Biber et al., 1998 , p. 4)

Por criteriosa entende-se que deva ela refletir a variedade escolhida o mais fielmente possível. Além de ser compatível com os objetivos da pesquisa (Hasan, 1992), a escolha deve ser feita com cuidado, visando a incorporar somente aquele material necessário para representar a amostra que se deseje. Por exemplo, se se quiser construir um corpus geral de uma língua, deve-se fazer uma escolha a mais variada possível: ela deve incluir o maior número possível de registros encontrados na língua-alvo e cada registro, por sua vez, deve ter o maior número possível de exemplares. Se por outro lado se desejar um corpus de uma vari-

edade específica, deve-se ser o mais seletivo possível na escolha dos exemplares, para que os mesmos reflitam de fato a variedade escolhida, ou seja, para que não hajam vieses nem contaminações.

A definição que incorpora as características principais já mencionadas nas anteriores é:

> 'Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise' (Sanchez, 1995, pp. 8-9)

Esta definição é mais completa porque incorpora vários pontos importantes:

(a) A origem: Os dados devem ser autênticos
(b) O propósito: O corpus deve ter a finalidade de ser um objeto de estudo lingüístico
(c) A composição: O conteúdo do corpus deve ser criteriosamente escolhido
(d) A formatação: Os dados do corpus devem ser legíveis por computador
(e) A representatividade: O corpus deve ser representativo de uma língua ou variedade
(f) A extensão: O corpus deve ser vasto para ser representativo

Em resumo, os quatro pré-requisitos para a formação de um corpus computadorizado são:

(1) Primeiramente, o corpus deve ser composto de textos autênticos, em linguagem natural. Assim, os textos não podem ter sido produzidos com o propósito de serem alvo de pesquisa lingüística. E não podem ter sido criados em linguagem artificial, tais como linguagem de programação de computadores ou notação matemática.

(2) Em segundo lugar, quando se fala em autenticidade dos textos, subentende-se textos escritos por falantes nativos. Tanto assim que, quando este não é o caso, deve-se qualificá-lo, falando-se em corpora 'de aprendizes' ('learner corpora').

(3) O terceiro pré-requisito é que o conteúdo do corpus seja escolhido criteriosamente. Os princípios da escolha dos textos devem seguir, acima de tudo, as condições de naturalidade e autenticidade. Mas devem também obedecer a um conjunto de regras estabelecidas pelos seus criadores de modo que o corpus coletado corresponda às características que se deseja dele. Ou seja, o conteúdo do corpus deve ser selecionado a fim de garantir que o corpus tenha uma certa característica. Por exemplo, se o desejo é construir um corpus de português brasileiro escrito que represente a língua portuguesa, tal qual ela é escrita no Brasil, em sua totalidade, a coleta deve ser guiada por um conjunto de critérios que garanta, entre outras coisas, que o maior número possível de tipos textuais existentes no português brasileiro esteja representado, que haja uma quantidade aceitável de cada tipo de texto e que a seleção dos textos seja aleatória, a fim de que não se contamine a coleta com variáveis indesejáveis.

(4) O quarto pré-requisito é mais problemático: representatividade. Tradicionalmente, tende-se a ver um corpus como um conjunto representativo de uma variedade lingüística ou mesmo de um idioma. Mas a questão não pode ser enfocada no vácuo. Cabe se perguntar 'representativo do quê?' e 'representativo para quem?'. A questão da representatividade é discutida abaixo com mais detalhes.

3.1. Tipologia

A nomenclatura empregada na Lingüística de Corpus para se definir o conteúdo e o propósito dos corpora é muito extensa. Os tipos principais citados na literatura são apresentados abaixo, agrupados segundo alguns critérios:

Modo

- Falado: Composto de porções de fala transcritas.
- Escrito: Composto de textos escritos, impressos ou não.

Tempo

- Sincrônico: Compreende um período de tempo.
- Diacrônico: Compreende vários períodos de tempo.
- Contemporâneo: Representa o período de tempo corrente.
- Histórico: Representa um período de tempo passado.

Seleção

- De amostragem (*sample corpus*): Composto por porções de textos ou de variedades textuais, planejado para ser uma amostra finita da linguagem como um todo.
- Monitor: A composição é reciclada para refletir o estado atual de uma língua. Opõe-se a corpora de amostragem.
- Dinâmico ou orgânico: O crescimento e diminuição são permitidos, qualifica o corpus monitor.
- Estático: Oposto de dinâmico, caracteriza o corpus de amostragem.
- Equilibrado (*balanced*): Os componentes (gêneros, textos, etc.) são distribuídos em quantidades semelhantes (por exemplo, mesmo número de textos por gênero).

Conteúdo

- Especializado: Os textos são de tipos específicos (em geral gêneros ou registros definidos).
- Regional ou dialetal: Os textos são provenientes de uma ou mais variedades sociolingüísticas específicas.
- Multilíngüe: Inclui idiomas diferentes.

Autoria

- De aprendiz: Os autores dos textos não são falantes nativos.
- De língua nativa: Os autores são falantes nativos.

Disposição interna

- Paralelo: Os textos são comparáveis (p.ex. original e tradução).
- Alinhado: As traduções aparecem abaixo de cada linha do original.

Finalidade

- De estudo: O corpus que se pretende descrever.
- De referência: Usado para fins de contraste com o corpus de estudo.
- De treinamento ou teste: Construído para permitir o desenvolvimento de aplicações e ferramentas de análise.

Além dos critérios acima, é possível propor alguns outros meios para a classificação dos corpora segundo sua composição:

(a) Pluralidade de autoria: Os textos[3] foram produzidos por um autor apenas ou mais?
(b) Origem da autoria: Os textos foram produzidos por falantes nativos ou não-nativos?
(c) Meio: Os textos foram escritos ou falados[4] ?
(d) Integralidade: Os elementos do corpus são textos integrais ou fragmentos?
(e) Especificidade: O corpus é composto de tipos variados de texto ou textos específicos?
(f) Dialeto: As variedades presentes no corpus são do tipo 'padrão' ou regionais / dialetais?
(g) Equilíbrio: As variedades do corpus são distribuídas eqüitativamente ou não?
(h) Fechamento: É permitida a inclusão de conteúdos novos ou não?
(i) Renovação: O conteúdo do corpus reflete um período definitivo de tempo ou se renova?
(j) Temporalidade: O corpus é planejado para retratar períodos históricos de tempo ou não?

[3] Aqui entende-se por texto uma amostra de linguagem falada ou escrita delimitada segundo critérios dos compiladores do corpus.

[4] Embora incomum, pode-se afinar esta classificação diferenciando-se textos escritos para serem lidos de textos escritos para serem falados (roteiros, palestras, etc), e textos falados para serem ouvidos de textos falados para serem escritos (isto é, ditados).

(k) Plurilingüísmo: O corpus possui só textos originais ou também as traduções destes textos para uma ou mais línguas?
• Intercalação: As traduções dos textos são incorporadas a cada linha do texto original ou vêm em textos separados?
(A partir de Atkins et al., 1992, p. 6):

3.2. Representatividade

Na sua essência, um corpus, seja de que tipo for, é tido como representativo da linguagem, de um idioma, ou de uma variedade dele. Ou, como diz Leech, o corpus possui uma *função representativa*. A característica mais facilmente associada à representatividade é justamente a extensão do corpus, o que significa em termos simples que para ter representatividade o corpus deve ser o maior possível (Sinclair, 1991; vide seção a seguir). Isto se deve a dois fatores:

(a) A linguagem é um sistema probabilístico (Halliday, 1991, 1992), onde certos traços são mais freqüentes que outros:

(i) No caso do léxico, pode-se diferenciar as palavras entre aquelas de 'maior freqüência' e as de 'menor freqüência', sendo que a diferença entre elas é relativa. Assim, algumas palavras têm freqüência de ocorrência muito rara e, para que haja probabilidade de ocorrerem no corpus, é necessário incorporar-se uma quantidade grande de palavras ao corpus. Em outras palavras, quanto maior a quantidade de palavras, mais probabilidade há de palavras de baixa freqüência aparecerem.

(ii) No caso dos sentidos das palavras, também se pode distinguir entre os sentidos mais freqüentes e os menos freqüentes dos itens lexicais. Assim, mesmo palavras de alta freqüência têm sentidos raros (por exemplo, 'serviço' entendido como 'saque' no jogo de tênis) e, portanto, esses sentidos terão maior probabilidade de ocorrer quanto maior for o corpus.

(b) O corpus é uma amostra de uma população cuja dimensão não se conhece (a linguagem como um todo). Desse modo, não se pode estabelecer qual seria o tamanho ideal da amostra para que ela represente esta população. Uma salvaguarda neste caso é tornar a

amostra a maior possível (Sinclair, 1991), a fim de que ela se aproxime o mais possível da população da qual deriva, sendo assim mais representativa.

Não há critérios objetivos para a determinação da representatividade. Por isso, uma amostra deve ter, além das características acima mencionadas, uma dada extensão. Assim, quando se diz que um corpus deva ser representativo, entende-se representatividade em termos da extensão do corpus, isto é, de um número determinado de palavras e de textos. Isso suscita de imediato duas questões: representativo do quê? Para quem?

Para se responder à questão 'representativo do quê?', deve-se olhar para a questão da amostragem. Para que qualquer amostra seja representativa, é necessário se conhecer a população da qual ela provém. No caso da linguagem, a dimensão da população total é desconhecida. Por isso, não é possível estimar-se qual seria uma amostra representativa da linguagem e, portanto, estritamente falando, não se pode afirmar que um corpus qualquer seja representativo.

Embora não se possa falar em representatividade em termos absolutos, pode-se tratar da questão em termos relativos. A principal maneira, ou 'salvaguarda' (Sinclair, 1991), pela qual se pode garantir maior representatividade é através do aumento da extensão do corpus. Um corpus maior é em geral mais representativo do que um menor devido ao fato de conter mais instâncias de traços lingüísticos raros.

A representatividade está ligada à questão da probabilidade. A linguagem é de caráter probabilístico (vide acima), e, sendo assim, há a possibilidade de estabelecer uma relação entre traços que são mais comuns e menos comuns em determinado contexto. O conhecimento da probabilidade de ocorrência de traços lexicais, estruturais, pragmáticos, discursivos, etc. está no cerne da Lingüística de Corpus e, portanto, o conhecimento acerca da probabilidade de ocorrência da maioria dos traços lingüísticos em vários contextos ainda está sendo adquirido.

O campo do léxico, entretanto, é onde se possui a maior quantidade de conhecimento derivado do exame de corpora. Para esta discussão,

é necessário distinguir-se entre a forma e o sentido lexical. Em qualquer corpus, as formas de freqüência 1 (também conhecidas como 'hapax legomena') são a maioria. Baseando-se neste fato, é possível afirmar que o léxico de freqüência baixa é o mais comum, isto é, que a maioria das palavras de uma língua é composta de palavras que ocorrem poucas vezes. Em outras palavras, palavras de baixa freqüência têm uma probabilidade baixa de ocorrência (1 em 1 milhão, por exemplo) e, já que elas formam a maior parte do vocabulário de uma língua, é necessário usar amostras grandes para que tais palavras possam ocorrer.

O sentido das palavras também entra em jogo na discussão da representatividade. A freqüência das formas em si não é suficiente, porque mesmo palavras de alta freqüência possuem vários sentidos. Assim, uma freqüência alta pode 'esconder' vários sentidos, os quais separados teriam baixa freqüência. Para que seja representativo, um corpus deve conter o maior número possível de sentidos de cada forma. Por exemplo, a forma 'como' pode significar a preposição ou a primeira pessoa do singular do verbo comer no presente do indicativo. Esta forma é comum na língua portuguesa, ocorrendo aproximadamente 531 vezes por milhão. Simplesmente olhando-se para a forma 'como' na listagem de freqüências do corpus não é possível se saber se ambos os sentidos estão representados. Um corpus geral que vise a representar a língua portuguesa deve conter ambos os sentidos deste vocábulo, já que ambas as formas existem na língua.

A extensão do corpus comporta três dimensões. A primeira é o número de palavras. O número de palavras é uma medida da representatividade do corpus no sentido de que quanto maior o número de palavras maior será a chance do corpus conter palavras de baixa freqüência, as quais formam a maioria das palavras de uma língua. A segunda é o número de textos, a qual se aplica a corpora de textos específicos. Um número de textos maior garante que este tipo textual, gênero, ou registro, esteja mais adequadamente representado. A terceira é o número de gêneros, registros ou tipos textuais. Esta dimensão se aplica a corpora variados, criados para representar uma língua como um todo. Aqui, um número maior de textos de vários tipos permite uma maior abrangência do espectro genérico da língua.

A outra perspectiva, a partir da qual se pode enfocar a questão da representatividade, é através da pergunta 'representativo para quem?'. Esta pergunta tem validade porque, conforme discutido acima, não se pode demonstrar, neste estágio do nosso conhecimento dos fenômenos de larga escala da linguagem, qual seria uma amostra representativa. Devido a isso, tem-se falado em representatividade como um 'ato de fé' (Leech, 1991, p.27). Em outras palavras, os usuários de um corpus atribuem a ele a função de serem representativos de uma certa variedade. O ônus é dos usuários em demonstrar a representatividade da amostra e de serem cuidadosos em relação à generalização dos seus achados para uma população inteira (um gênero ou a língua inteira, por exemplo).

Um grande problema é que a quantidade mínima de dados necessários para a formação de um corpus nunca foi estimada (Berber Sardinha, no prelo), sendo o critério de tamanho empregado subjetivamente na definição de corpus. Este é o tema da próxima seção.

3.3. Extensão

Embora seja um critério fundamental na representatividade, pouco se tem pesquisado a questão da definição de critérios mínimos de extensão para a constituição de um corpus representativo. Pode-se definir três abordagens:

- Impressionística: baseia-se em constatações derivadas da prática da criação e da exploração de corpora, em geral feita por autoridades da área. Por exemplo, Aston (1997) menciona patamares que caracterizariam um corpus pequeno (20 a 200 mil palavras) e um grande (100 milhões ou mais). Leech (1991) fala de 1 milhão de palavras como a taxa usual ('going rate'), sugerindo o que seja o patamar mínimo. Outros são mais vagos, como Sinclair (1996), o qual postula que o corpus deva ser tão grande quanto a tecnologia permitir para a época, deixando-se subentender que a extensão de um corpus deva variar de acordo com o padrão corrente nos grandes centros de pesquisa, que possuem equipamentos de última geração.

• Histórica: fundamenta-se na monitoração dos corpora efetivamente usados pela comunidade. Por exemplo, Berber Sardinha (no prelo) sugere uma classificação baseada na observação dos corpora utilizados, segundo quatro anos de conferências de Lingüística de Corpus:

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Graficamente, a escala seria esta:

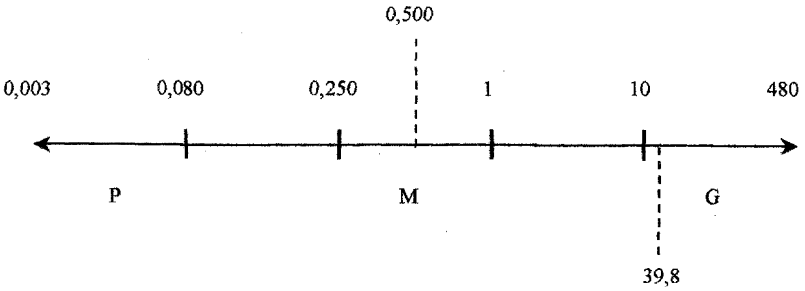

Figura 1: Escala de tamanho relativo de corpora. Os números referem-se a quantias em milhões, e as letras a 'Pequeno', 'Médio', e Grande'. O número sobre a linha tracejada superior indica a mediana e, sob a inferior, a média aritmética.

• Estatística: fundamenta-se na aplicação de teorias estatísticas. Por exemplo, Biber (1993) emprega fórmulas matemáticas para identificar quantidades mínimas de palavras, gêneros e textos que se constituiriam em uma amostra representativa. Pode ser subdividida em três vertentes:

(1) Interna: Dado um corpus pré-existente que serve como amostra maior, qual o tamanho mínimo de uma amostra que mantém estáveis as características desta amostra maior? Esta é a perspectiva seguida por Biber (1990, 1993).

(2) Externa: Dada uma fonte externa de referência cuja dimensão é conhecida, qual o tamanho do corpus necessário para representar majoritariamente esta fonte? Esta vertente tem sido discutida pela comunidade de lingüistas do corpus (Berber Sardinha, 1998).

(3) Relativa: Quanto se perderia se o corpus fosse de um tamanho x? Dados meus recursos existentes, quais parâmetros posso utilizar para abalizar minha decisão relativa ao tamanho de corpus que posso compilar? Uma proposta, segundo esta perspectiva ainda não foi formalizada, mas está presente, por exemplo, em Sanchez e Cantos (1997a, b), os quais estimam matematicamente a quantidade do vocabulário presente em corpora de diversos tamanhos hipotéticos. Uma proposta similar é apresentada por Yang e Song (1998), os quais fazem uma previsão da quantidade de dados necessários para incluir certas características gramaticais.

3.3.1. Especificidade

Um modo de atingir a representatividade total de um corpus é incluir nele toda a linguagem. Como isto é impossível para um idioma inteiro, a possibilidade mais próxima é restringir o conteúdo a, por exemplo, um autor apenas. Assim, a coletânea de todos os trabalhos escritos por Shakespeare seria um corpus representativo deste autor. Uma outra maneira é delimitar, ao máximo, a variedade (tipo de texto, por exemplo) incluída no corpus. Isto ocorre porque uma variedade específica da linguagem demonstra uma maior padronização e conseqüente menor variação no nível do léxico, gramática, discurso, etc, Ou seja, apresenta maior grau de 'fechamento' (*closure*) (McEnery e Wilson, 1996).

Os corpora gerais podem ser usados, obviamente, como fonte para criação de corpora especializados. O British National Corpus, por exemplo, possui uma quantidade grande de artigos de pesquisa e, portanto, o

usuário pode extrair estes textos e criar um sub-corpus especializado de artigos científicos. A vantagem de se aproveitar os recursos de grandes corpora neste sentido é, evidentemente, que o usuário não necessitará coletar um corpus novo. Além disso, no caso do BNC, o usuário já disporá de textos anotados e etiquetados gramaticalmente, o que novamente lhe poupará tempo e recursos.

Entretanto, a quantidade de textos de uma variedade ou domínio específicos nos corpora gerais é pequena. Assim, normalmente, corpora compilados em pequena escala por pesquisadores individuais acabam sendo mais representativos do que os respectivos sub-corpora dos corpora gerais. Aston (1997), por exemplo, mostra que o seu corpus de artigos acadêmicos de pesquisa sobre hepatite C é mais completo e representativo do que um equivalente extraído do BNC.

Um problema com muitos corpora específicos é que eles são geralmente criados com o propósito de servirem a projetos particulares e, por várias razões (direitos autorais, inclusive), não são colocados à disposição da comunidade científica. Desse modo, não satisfazem a condição de serem dados verificáveis, o que compromete a pesquisa em termos de sua replicabilidade e generabilidade.

3.3.2. Adequação

Um outro critério fundamental na composição de um corpus é a adequação. Este aspecto envolve os criadores do corpus, mas atinge principalmente os seus usuários. Por mais que muitos dos corpora tentem ser representativos de uma língua como um todo ou de uma variedade dela, eles não são necessariamente adequados à investigação de qualquer característica lingüística. Conforme lembra Hasan:

> 'Para serem adequados, os corpora devem ser afinados com os objetivos da análise. Suponha que meu interesse seja em perguntar: Qual a freqüência do sujeito pronominal em inglês? É possível que 22 mil orações possam se constituir em evidência adequada. Mas dado o meu interesse em analisar os dados num certo grau de delicadeza, (…) eu precisaria de um corpus muito maior.' (Hasan, 1992, p. 301)

Em outras palavras, embora representativo, o corpus possui seus limites. Ele pode ajudar a responder apenas alguns tipos de perguntas. Com esta postura, parte-se da pesquisa e não do objeto. Ou seja, invertendo-se a origem da empreitada, coloca-se a questão de pesquisa na frente do objeto. Além de representativo, o corpus deve ser adequado aos interesses do pesquisador. Quer dizer, em vez de se dizer, 'eu tenho este corpus, então agora vou descrevê-lo', deve-se pensar 'eu desejo investigar esta questão, então eu necessito de um corpus com estas características'.

A adequação do corpus é tomada como dada. Assume-se que o corpus com o qual se esteja lidando e as perguntas que se faz a ela sejam adequadas para os propósitos da investigação. Sem isso, a pesquisa perde o sentido.

A colocação da adequação do corpus, antes de tudo, na pesquisa em Lingüística de Corpus, tem como conseqüência o questionamento da validade de corpora gerais. Tais corpora têm sido a norma na área e incluem os célebres Brown, LOB, London Lund e BNC. Eles foram construídos com o intuito de servirem como representantes de uma língua como um todo, ou mais especificamente de um dialeto ou variante. Por exemplo, o corpus Brown tem sido tido por muito tempo como representante do inglês americano escrito. O LOB, por sua vez, é tido como representante do inglês britânico escrito. O London-Lund é considerado representativo do inglês britânico falado. Finalmente, o BNC é o mais ambicioso, pois é tido como representante do inglês britânico, tanto do modo falado quanto escrito. Uma característica importante dos corpora citados aqui é que eles são disponibilizados para a comunidade acadêmica e, assim, cumprem seu papel de fontes de dados verificáveis.

4. Teorias de linguagem e Lingüística de Corpus

A Lingüística de Corpus trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem enquanto sistema probabilístico. O empirismo é, em termos bem simples, uma doutrina filosófica segundo a qual o conhecimento se origina da

experiência. Na lingüística, o empirismo significa dar primazia aos dados provenientes da observação da linguagem, em geral reunidos sob a forma de um corpus. O empirismo se coloca em oposição ao racionalismo, segundo o qual, em linhas gerais, o conhecimento provém de princípios, estabelecidos *a priori.* O racionalismo, na lingüística, se fundamenta no estudo da linguagem através da introspecção, como meio de verificar modelos de funcionamento estrutural e processamento cognitivo da linguagem. Há, portanto, uma oposição fundamental entre as posições filosóficas inerentes às visões empirista e racionalista da linguagem, expressas por meio dos programas de pesquisa de seus maiores expoentes. De um lado, Halliday, seguindo a tradição empirista, e de outro Chomsky, o maior expoente do racionalismo na lingüística.

O segundo elemento central da conceituação em que a Lingüística de Corpus se baseia é a visão probabilística da linguagem. Aqui fica mais evidente a oposição entre Halliday e Chomsky. Halliday vê a linguagem como *probabilidade*, enquanto Chomsky a enxerga como *possibilidade* (Kennedy, 1998). A lingüística Chomskyana gerativista enfatiza a determinação de quais agrupamentos sintáticos são possíveis (i.e. permissíveis) dado o conhecimento que um falante nativo possui de sua língua. Já a lingüística Hallidayana descreve a probabilidade dos sistemas lingüísticos, dados os contextos em que os falantes os empregam.

A visão da linguagem enquanto sistema probabilístico pressupõe que embora muitos traços lingüísticos sejam possíveis teoricamente, eles não ocorrem com a mesma freqüência. Podemos citar dois exemplos. Primeiramente, no nível morfossintático, a freqüência de substantivos (no inglês e, com certeza, no português) é maior do que qualquer outra categoria; cerca de 25% das palavras são substantivos (Kennedy, 1998, p.103). Desse modo, a probabilidade de um traço ser um substantivo é maior do que outra classe gramatical. E em segundo lugar, embora seja teoricamente possível se aninhar orações relativas *ad infinitum* (o gato que está no tapete é meu, o gato que está no tapete que é meu é pardo, o gato que está no tapete que é meu que é pardo está dormindo, etc), à primeira vista a freqüência de ocorrência de frases com mais de uma oração relativa é muito maior do que com sucessivas orações. Em resumo, as possibilidades da estrutura não se realizam todas com a mesma freqüência.

O mais importante da diferença de freqüências entre os traços é o fato de essas diferenças não serem aleatórias. Se o fossem, então o fato das possibilidades estruturais se realizarem com freqüências diferentes não seria significativo, isto é, não acrescentaria informação a respeito da própria estrutura. Entretanto, pelo contrário, há um mapeamento regular entre a freqüência maior ou menor de um traço e um contexto de ocorrência. Ou, nas palavras de Biber (1988, 1995), há uma correlação entre características lingüísticas e situacionais (os contextos de uso). O conjunto da pesquisa desenvolvida por Biber apresenta evidências inequívocas de que conjuntos de traços lingüísticos variam sistematicamente com relação a textos típicos de contextos comunicativos específicos. Em outras palavras, a variação não é aleatória.

Quando se diz que a variação não é aleatória, na verdade, está se afirmando que a linguagem é *padronizada* ('patterned'). A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura, que se repete significativamente, mostra sinais de ser na verdade um *padrão* lexical ou léxico-gramatical. A linguagem forma padrões que apresentam regularidade (se mostram estáveis em momentos distintos, isto é, tem freqüência comparável em corpora distintos) e variação sistemática (correlacionam-se com variedades textuais, genéricas, dialetais, etc). Exemplos notáveis da descrição da linguagem por meio da indução de padrões recorrentes são a gramática de verbos (Francis, G. e Hunston, 1996) e de substantivos e adjetivos (Francis, G. e Hunston, 1998) lançadas pelo projeto COBUILD[5], nas quais se descreve exaustivamente todos os padrões lexicais existentes na língua inglesa.

Por isso, além da possibilidade teórica de ocorrência, uma teoria da linguagem deve incorporar a probabilidade de ocorrência dos traços. Aqui a lingüística Chomskyana recorre à introspecção, ou à intuição do falante nativo, para responder a esta questão. Entretanto, o que o falante nativo pode informar é somente se o traço ou estrutura em questão é *intuitivamente* provável ou não, pois:

[5] Para uma crítica da descrição gramatical nesta linha vide Owen (1992).

> 'o ser humano, ao contrário do que em geral se pensa, não é bem organizado para isolar conscientemente o que é central e típico da linguagem; aquilo que é incomum é percebido imediatamente, mas os eventos costumeiros do dia-a-dia são apreciados subliminarmente.' (Sinclair e Renouf, 1988, p.151, tradução minha)

Para se saber qual a probabilidade de um traço ou estrutura é necessária, portanto, a observação empírica da freqüência do emprego, realizado por diversos usuários, em contextos definidos.

Destas considerações, tira-se duas conclusões. A primeira é a importância primordial de um corpus como fonte de informação, pois ele registra a linguagem natural realmente utilizada por falantes e escritores da língua em situações reais. A segunda é a não-trivialidade da investigação da freqüência de ocorrência de traços lingüísticos de várias ordens (lexicais, sintáticos, semânticos, discursivos, etc), pois é através do conhecimento da freqüência atestada que se pode estimar a probabilidade teórica.

Chomsky ridicularizou esta postura com sua famosa frase 'I live in Dayton, Ohio', empregada por ele em uma palestra no final dos anos 50. Ele argumentava que embora esta frase seja menos freqüente que 'I live in New York' (já que há mais pessoas em Nova York do que em Dayton), a diferença de freqüência de uso é totalmente irrelevante para uma teoria da linguagem, já que é ocasionada por uma realidade demográfica. Em primeiro lugar, a suposição de que 'I live in New York' é mais freqüente é somente isto, uma suposição. De fato não sabemos se os falantes da cidade usam esta frase e, se o fazem, em quais ocasiões e com qual freqüência. E, em segundo lugar, se descobrirmos, após um levantamento baseado em freqüências atestadas em um corpus, que os contextos em que se refere o local onde se mora apresenta-se desta forma e não de outras maneiras equivalentes ('I live in Manhattan', 'in this city', 'NYC', etc), teremos na verdade descoberto fatos sobre a linguagem até então desconhecidos. O conhecimento obtido não seria de modo algum trivial, pois nos informaria, entre outras coisas, como dizer o local de nossa moradia da maneira mais aceitável dentro de cada situação em que temos de fornecer esta informação, como grupos de falantes diferentes expressam-se em face de demandas lingüísticas similares, etc.

Uma teoria da linguagem torna-se mais pobre e ineficiente, ao não levar em conta estes dados.

Pode-se resumir através das seguintes características as diferenças entre a Lingüística de Corpus e a lingüística Chomskyana:

(a) Foco no desempenho lingüístico, em vez de competência;
(b) Foco na descrição lingüística, em vez de universais lingüísticos;
(c) Foco numa visão mais empirista do que racionalista da pesquisa científica (Leech, 1992, p.107, tradução minha).

Os modelos estruturais da linguagem em geral (incluindo os gerativistas de Chomsky) descrevem a linguagem através de esquemas 'slot and filler', nos quais as lacunas ('slots') sintáticas podem ser preenchidas lexicalmente de qualquer modo, desde que o conjunto de lacunas seja estruturalmente plausível. Esta visão tem críticos ferozes dentro da Lingüística de Corpus, dentre os quais destaca-se John Sinclair. O programa de pesquisa de Sinclair tem se pautado pela descrição da linguagem do ponto de vista lexical, cuja perspectiva é a descrição de quais agrupamentos lexicais são realmente empregados pelos falantes, isto é, atestados pelo uso. Esta perspectiva se concretizou em um princípio de entendimento da linguagem chamado de 'idiomático' (*idiom principle*), explicado como o fato do usuário de uma língua ter à sua disposição 'um grande número de frases pré– ou semi-construídas, que se constituem em escolhas únicas, muito embora pareçam analisáveis em segmentos' (Sinclair, 1987, p. 320, tradução minha).

Esta visão da linguagem enquanto sendo formada por porções lexicais (*chunks*) ou idiomas é compartilhado por outros autores trabalhando em contextos diferentes (e.g. Bolinger, 1976; Nattinger e DeCarrico, 1992; Pawley e Syder, 1983). Notadamente, Pawley e Syder (1983) foram influentes na ligação entre a presença de idiomas ou 'multipalavras' (multi-words) e a naturalidade da linguagem. Para eles, a 'naturalidade' e a percepção da 'fluência' na produção do falante nativo devem-se em boa medida ao emprego de um grande número de expressões pré-fabricadas e à união destas em seqüências maiores. Com base neste princípio, Nattinger e De Carrico (1992) produziram um levan-

tamento de frases idiomáticas, visando ao ensino e à aprendizagem de línguas. Mas foi a formulação de Sinclair (e a metodologia computacional desenvolvida por ele) que influenciou um grande número de trabalhos voltados nesta área. Além disso, o florescimento em geral da *fraseologia baseada em corpus* (e.g. Cowie, 1998; Moon, 1998) também deve muito ao trabalho pioneiro de Sinclair.

Haveria, segundo Sinclair, um espaço comum formado pelo léxico e pela sintaxe, no qual ambos são co-selecionados: a escolha de cada item lexical implica na redução das escolhas dos itens lexicais e das categorias gramaticais que podem segui-lo. Complementarmente, a escolha de uma classe gramatical reduz a escolha possível de classes gramaticais e de itens lexicais que podem seguir-se a ela. Já é possível descrever-se com muita precisão as probabilidades de certos itens ocorrerem em co-textos específicos, e, desse modo, os níveis do léxico e da gramática tornam-se supérfluos. Neste nível, a separação entre léxico e sintaxe é uma questão de conveniência analítica, sem respaldo empírico.

Uma teoria que admite este espaço é justamente a lingüística sistêmico-funcional de Halliday, na qual este nível é conhecido por léxico-gramática. Longe de ser uma coincidência, isto mostra a ligação íntima entre a perspectiva seguida pela Lingüística de Corpus e pela lingüística Hallidayana.

A conexão existe porque Halliday é um exemplo de lingüista de inclinação empirista, entretanto ele não é (i.e. não se auto-define como) um lingüista do corpus. A formulação das teorias de Halliday, na forma da lingüística sistêmico-funcional, não se pauta pela exigência de um corpus nem do instrumental comumente empregado pelos lingüistas do corpus. Entretanto, a sua visão de linguagem se encaixa perfeitamente nos preceitos da Lingüística de Corpus e serve como arcabouço teórico maior no qual ela se pode incluir.

Um lingüista que critica a posição de antagonismo entre lingüistas do corpus e os demais, traçada nesta seção, é Charles Fillmore. Ele faz uma caricatura dos dois tipos de lingüista. Segundo ele, o lingüista de corpus seria aquele que 'possui todos os fatos primários que necessita, na forma de um corpus de aproximadamente um zilhão de palavras' e

que se dedica a 'derivar fatos secundários a partir de fatos primários'. O outro tipo de lingüista é chamado por Fillmore de 'lingüista de poltrona' e demonstraria o seguinte comportamento:

> "Ele se senta numa poltrona bem confortável, com os olhos fechados e com a cabeça apoiada nas mãos por trás. De vez em quando ele abre os olhos, se mexe todo, berra 'Nossa, que fato interessante!', pega o lápis e toma algumas notas (…) ficando entusiasmado por ter chegado mais perto de entender como a linguagem funciona." (Fillmore, 1992, p. 35)

Quando se encontram, os dois lingüistas se estranham – o de poltrona indaga 'por que eu deveria acreditar que o que você me diz é interessante?', ao que o do corpus retruca 'por que eu deveria acreditar que o que você me diz é verdadeiro?'. Embora o diálogo entre os dois seja difícil, para Fillmore, os dois lingüistas deveriam existir em harmonia na mesma pessoa, já que ambos tem a aprender com o outro.

5. Estatuto da Lingüística de Corpus

Um debate que se desenrola entre os praticantes da Lingüística de Corpus se centra na definição do status da área: é a Lingüística de Corpus uma disciplina ou metodologia? Claramente, a Lingüística de Corpus não é uma disciplina tal qual a psicolingüística, sociolingüística, semântica, etc., pois seu objeto de pesquisa não é delimitado como em outras áreas. A Lingüística de Corpus não se dedica a um assunto definido (Leech, 1992, p.106). Pelo contrário, ela se ocupa de vários fenômenos comumente enfocados em outras áreas (léxico, sintaxe, textura, etc.). Seria então seguro se concluir que a Lingüística de Corpus é então uma metodologia da qual outras áreas podem se fazer valer? A princípio sim. McEnery e Wilson (1996), por exemplo, afirmam que a Lingüística de Corpus é 'apenas uma metodologia' (p.1), e Leech (1992, p. 105) a descreve como uma 'base metodológica'.

Mas se a Lingüística de Corpus é metodologia ou não, vai depender da definição de metodologia que se está usando. Se entendermos metodologia como *instrumental*, então é possível aplicar-se o instru-

mental da Lingüística de Corpus livremente e manter a orientação teórica da disciplina original. Desse modo, teríamos, por exemplo, a sintaxe baseada em corpus *versus* a sintaxe 'tradicional', a fonologia baseada em corpus *versus* a fonologia 'tradicional' e assim por diante. Tudo o que mudaria entre estas vertentes opostas seria o instrumental; os dados, a orientação, os pressupostos teóricos, as implicações dos resultados e tudo o mais permaneceria o mesmo.

Mas a Lingüística de Corpus não se resume a um conjunto de ferramentas. Assim, se entendermos metodologia como um *modo típico de aplicar um conjunto de pressupostos de caráter teórico*, então a Lingüística de Corpus pode ser entendida como uma metodologia, pois traz consigo algo mais do que simplesmente o instrumental computacional. Aqui se encaixam as investigações do comportamento do léxico, típicas de lingüistas do corpus auto-definidos, como John Sinclair. A pesquisa de Sinclair acerca da colocação entre itens lexicais, por exemplo, não encontra espaço em outras disciplinas. Ela possui caráter essencialmente ascendente e tem como doutrina a não categorização a priori ('trust the text' é o seu lema). Por isso, exemplifica com precisão a prática empirista e situa-se como o pólo mais distante das abordagens racionalistas. Aliás, foi por isso mesmo que uma das maiores correntes de pesquisa em corpus surgiu.

Uma outra razão pela qual a Lingüística de Corpus não é uma metodologia é o fato de seus praticantes produzirem conhecimento novo, muito do qual é de caráter contestatório de práticas e preceitos correntes:

> 'Embora o escopo da Lingüística de Corpus possa ser definido em termos do que as pessoas fazem com corpora, seria um engano assumir que Lingüística de Corpus é somente um meio mais rápido de descrever como a linguagem funciona (…) A análise de um corpus pode revelar, e freqüentemente revela, fatos a respeito de uma língua que nunca se pensou em procurar.' (Kennedy, 1998, p. 9, tradução minha)

O exemplo mais imediato é a contestação dos pressupostos da lingüística gerativa, delineada acima.

O fato de a Lingüística de Corpus produzir conhecimento de natureza distinta e até contestatória a coloca de certo modo em condições

similares à Lingüística Aplicada. A Lingüística Aplicada não é mais vista como um simples espaço no qual se aplicam os conhecimentos produzidos na lingüística. Os conhecimentos que se aplicam na lingüística Aplicada não são de origem exclusiva da lingüística. Por isso ela possui um caráter essencialmente transdisciplinar (cf. Celani, 1998).

Uma terceira possibilidade que se apresenta é que a Lingüística de Corpus não é nem disciplina nem metodologia. Segundo Hoey:

> 'Lingüística de Corpus não é um ramo da lingüística, mas a rota para a lingüística' (Hoey, 1997, tradução minha)

Esta definição se assemelha a dizer que a Lingüística de Corpus é uma perspectiva, isto é, uma maneira de se chegar à linguagem. Esta definição faz alusão ao conceito de teoria lingüística enquanto 'janela' que molda como enxergamos a linguagem (Pike, 1972). Dessa forma, segundo Hoey (1997) a Lingüística de Corpus não seria apenas um instrumental, mas sim um *abordagem*. De modo similar, Leech (1992, p.106) a define como:

> 'A Lingüística de Corpus define não somente uma nova metodologia emergente para o estudo da linguagem, mas uma nova empreitada de pesquisa e, na verdade, uma nova abordagem filosófica.'

Daí a preferência de alguns influentes lingüistas do corpus, como Douglas Biber, pelo termo 'abordagem baseada em corpus'. Tanto assim que em seu livro mais recente, o título é 'Corpus Linguistics', mas esta expressão mais conhecida só aparece na capa, sendo substituída por 'corpus-based approach' no decorrer da obra.

6. Tipos de pesquisa privilegiadas

Há uma quantidade enorme de trabalhos que se encaixam na Lingüística de Corpus e o número cresce a cada ano. Segundo McEnery e Wilson (1996, p.18) teriam aparecido 620 trabalhos em 25 anos de ati-

vidade (de 1965 a 1991), mas quase a metade teria surgido nos últimos cinco anos apenas. A despeito de sua diversidade, os trabalhos em Lingüística de Corpus compartilham de algumas características em comum:

(a) São empíricos e analisam os padrões reais de uso em textos naturais.
(b) Utilizam coletâneas grandes e criteriosas de textos naturais, conhecidas por 'corpus', como a base da análise.
(c) Fazem uso extensivo de computadores na análise, empregando técnicas automáticas e interativas.
(d) Dependem de técnicas quantitativas e qualitativas. (Biber et al., 1998, p. 4)

Pode-se pensar em três paradigmas de pesquisa em Lingüística de Corpus que partilhariam em maior ou menor grau as características acima:

(1) Paradigma informal baseado em concordâncias

(2) Paradigma estatístico baseado em modelos *log-linear*

(3) Paradigma estatístico fundamentado em Modelos Ocultos de Markov (Leech, 1992, pp.114-120)

O paradigma que concentra a maior parte das pesquisas é o primeiro, que se pauta pela descrição da linguagem com pouco ou nenhum suporte estatístico. Os demais paradigmas assumem uma perspectiva quantitativa mais sólida e fazem uso de técnicas estatísticas mais avançadas.

Juntamente com a explosão do número de trabalhos em Lingüística de Corpus, há um crescimento de áreas de pesquisa privilegiadas. Kennedy (1998, p. 9), cita quatro concentrações principais:

(1) compilação de corpus

(2) desenvolvimento de ferramentas

(3) descrição da linguagem

(4) aplicação de corpora (ensino de línguas, reconhecimento de voz, tradução, etc)

A área na qual há mais atividade é a terceira, a da descrição. Há um número considerável de trabalhos que enfocam principalmente o léxico e a gramática a partir do exame de um corpus. Estes trabalhos se ocupam do que Kennedy (1991, p.98)chama de 'ecologia lingüística', isto é, do comportamento de itens lexicais ou de estruturas gramaticais no seu 'habitat' natural (o meio lingüístico que ocorrem).

As questões de que se ocupa a área da descrição são tipicamente as seguintes:

(a) Quais os padrões lexicais dos quais a palavra faz parte?
(b) A palavra se associa regularmente com outros sentidos específicos?
(c) Em quais estruturas ela aparece?
(d) Há uma correlação entre o uso/sentido da palavra e as estruturas das quais ela participa?
(e) A palavra está associada com (uma certa posição na) organização textual? (Hoey, 1997, p. 3)

A maioria destas questões centraliza-se na descrição de três fenômenos:

(1) *Colocação*: associação entre itens lexicais, ou entre o léxico e campos semânticos. Por exemplo, em termos lexicais, 'stark' associa-se a 'contrast'; 'sheer', a 'scale', 'number' e 'force' (Partington, 1998). Em termos de campos semânticos, 'jam' relaciona-se com itens do campo de 'alimentos': 'tarts', 'butty' e 'doughnuts' (Moon, 1998, p.27).

(2) *Coligação*: associação entre itens lexicais e gramaticais. Por exemplo, 'start' é mais comum com sintagmas nominais e orações –ing, enquanto 'begin' é mais usado com um complemento 'to'(Biber et al., 1998).

(3) *Prosódia semântica*: associação entre itens lexicais e conotação (negativa, positiva ou neutra) de campos semânticos. O nome deve-se ao fato de certas palavras prepararem o ouvinte ou o leitor para o conteúdo semântico que está por vir, da mesma maneira que a prosódia na fala indica para o interlocutor que tipos de sons estão por vir a seguir (Hoey, 1997, p.4). Por exemplo, 'cause' tem uma prosódia semântica

negativa, pois associa-se a palavras desfavoráveis como 'problem(s)', 'damage', 'death(s)', 'disease', 'concern' e 'cancer'. Já 'provide' possui uma prosódia semântica positiva ou neutra, já que se associa a palavras deste tipo, tais como 'assistance', 'care', 'jobs', 'opportunities' e 'training' (Stubbs, 1995).

O fenômeno da colocação é o mais tradicionalmente enfocado no estudo de corpus. Foi originalmente introduzido por Firth (1957) e explicado por sua famosa frase: 'you shall judge a word by the company it keeps'. Há três definições de colocação principais na literatura, segundo Partington (1998, pp. 16-17):

(1) Textual: 'Colocação é a ocorrência de duas ou mais palavras distantes um pequeno espaço de texto umas da outras' (Sinclair, 1991, p. 170)

(2) Psicológica: 'O sentido colocacional consiste das associações que uma palavra faz por conta dos sentidos das outras palavras que tendem a ocorrer no seu ambiente' (Leech, 1974, p. 20)

(3) Estatística: 'Colocação tem sido o nome dado à relação que um item lexical tem com itens que aparecem com probabilidade significativa no seu contexto (textual)' (Hoey, 1991, pp. 6-7)

Este elenco de questões se fundamenta na análise da palavra, pois segundo Hoey (1997), 'inevitavelmente se começa pela palavra'. Entretanto, esta é na verdade a abordagem baseada na *palavra*, na qual se privilegia o estudo da associação entre traços dentro de um pequeno cotexto (espaço de texto), geralmente quatro palavras para cada lado do item lexical de interesse. Esta é apenas um dos tipos de perspectivas possíveis da Lingüística de Corpus. A outra abordagem é a *textual* (Scott, 1997). Nesta perspectiva, o foco é a relação das palavras dentro do espaço do compreendido pelo texto inteiro. Segundo Scott (1997), este tipo de investigação da associação entre palavras captura com mais fidelidade o tipo de relação que Firth tinha em mente quando pensava em colocação. Firth ilustrava seu conceito com exemplos como 'letter' e 'postman', palavras estas que em geral não ocorrem dentro de um espaço estreito de poucas palavras, mas tendem a co-ocorrer em um mesmo

texto. A mudança de foco teria sido motivada pelas limitações tecnológicas da época (anos 60) em que se iniciou na prática, através do computador, a investigação da noção de colocação. Com os equipamentos da época, a computação da co-ocorrência lexical além de um certo espaço pequeno de texto, era inviável.

7. Comentários finais

O presente trabalho aproveitou o ensejo da comemoração dos 35 anos do corpus Brown para apresentar um painel do campo de investigação que esse corpus, em grande parte, ajudou a desenvolver: a lingüística de Corpus, que é hoje uma das áreas mais vibrantes voltadas ao estudo da linguagem. As dificuldades envolvidas em se retratar um campo tão vasto e dinâmico são pelo menos duas. A primeira, mais óbvia, refere-se à quantidade de trabalhos novos que surgem, muitos dos quais de grande impacto tanto na comunidade de lingüistas de corpus quanto fora dela. Um exemplo é a nova gramática da língua inglesa a ser lançada no final de 1999 por Douglas Biber e equipe (Biber et al., 1999). Essa gramática pretende ser a sucessora da célebre 'Comprehensive Grammar of the English Language' (Quirk et al., 1985) e foi feita totalmente a partir de um corpus de 40 milhões de palavras, cuidadosamente criado para esse fim. A nova gramática segue fielmente os modelos de descrição privilegiados pela Lingüística de Corpus (computação da freqüência, descrição da co-ocorrência e indução de padrões léxico-gramaticais) e desse modo contribuirá para que a descrição baseada em corpus se torne uma norma, em vez da exceção.

A segunda dificuldade refere-se ao debate de cunho teórico. A Lingüística de Corpus tem sido acusada de apenas fazer 'statement of facts', ou seja, de apenas registrar as ocorrências lexicais e estruturais. Para deixar de ser um tipo de 'Contabilidade lingüística', a Lingüística de Corpus necessita explicitar qual é o quadro teórico que lhe dá coerência e sustentação. Essa deficiência vem, em grande medida, do fato de os lingüistas de corpus não terem se preocupado com a plausibilidade psicológica (Leech, 1992, p.113) da área, ou seja, a Lingüística de Corpus 'ainda não considera central discutir o *porquê* da linguagem ser usada

de tal modo que ela exiba os padrões e fenômenos' que são retratados (Sch"nefeld, 1999, p.148; tradução minha, grifo no original). Neste sentido, a proposta da Psicologia Cognitiva de Langacker (1987, 2000) tem sido apresentada como alternativa para ajudar a preencher essa falta de explicação mental do fenômeno talvez mais central à Lingüística de Corpus, que é o da padronização.

Em suma, esses desenvolvimentos tanto descritivos quanto teóricos, prometem manter o dinamismo que caracterizará a Lingüística de Corpus por muitos anos ainda. Mas o real crescimento e pujança da área se manterá na medida em que mais e mais pesquisadores descubram no corpus uma fonte inestimável de informação, pois é no trabalho prático de exploração que a Lingüística de Corpus ganha vida (Leech, 1999). Assim, mais e mais estudantes, lingüistas e pesquisadores anônimos continuarão tendo a experiência de que fala Fillmore (1992, p. 35): 'não há nenhum corpus que contenha toda a informação que eu quero explorar', mas mesmo assim 'todo corpus me ensinou coisas sobre a linguagem que eu não teria descoberto de nenhum outro modo'.

Referências Bibliográficas

Aijmer, K. & B. Altenberg (Orgs.) (1991) *English Corpus Linguistics – Studies in honour of Jan Svartvik*. London: Longman.

Aston, G. (1997) Small and large corpora in language learning. Paper presented at the PALC Conference, University of Lodz, Poland, April 1997.

Atkins, S. et al. (1992) Corpus design criteria. *Literary and Linguistic Computing*, 7: 1-16.

Benson, M. et al (1986) *The BBI dictionary of English word combinations*. Amsterdam / Philadelphia: John Benjamins.

Berber Sardinha, A. P. (1998) Size of a representative corpus. Summary of discussion on CORPORA email discussion list, 26 August 1998.

_____. (1999) Processamento Computacional do Português. Simpósio, 9º. InPLA, PUCSP, São Paulo.

_____. (no prelo) O que é um corpus grande. *The ESPecialist*.

BIBER, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

_____. (1990) Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5: 257-269.

_____. (1993) Representativeness in corpus design. *Literary and Linguistic Computing*, 8: 243-257.

_____. (1995) *Dimensions of Register Variation – A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

BIBER, D. ET AL (1998) *Corpus linguistics – Investigating language structure and use*. Cambridge: Cambridge University Press.

_____. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

BIDERMAN, M. T. C. (1978) *Teoria Lingüística (Lingüística Quantitativa e Computacional)*. LTC: Rio de Janeiro / São Paulo.

BOLINGER, D. (1976) Meaning and memory. *Forum Linguisticum*, 1: 1-14.

CASTILHO, A. T. ET AL. (1995) Informatização de acervos da língua portuguesa. *Boletim da ABRALIN*, 17: 143-151. (Disponível na Internet: www.corpus.f2s.com/ataliba01.tif)

CELANI, M. A. A. (1998) Transdisciplinaridade na Linguistica Aplicada no Brasil. In: I. SIGNORINI & M. C. CAVALCANTI (org.). *Lingüística Aplicada e Transdisciplinaridade*. Campinas: Mercado de Letras.

CERMAK, F. (1997) Czech National Corpus: A case in many contexts. *International Journal of Corpus Linguistics*, 2.2: 181-198.

COWIE, A. P. (Org.) (1998) *Phraseology – Theory, Analysis, and Applications*. Oxford: Clarendon.

DUNCAN JR, J. C. (1972) *A Frequency Dictionary of Portuguese Words*. Tese de Ph.D., Stanford University.

FILLMORE, C. (1992) 'Corpus linguistics' or 'computer corpus linguistics'. In: J. SVARTVIK (org.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin, New York: De Gruyter.

FIRTH, J. R. (1957) *Papers in Linguistics – 1934-1951*. Oxford: Oxford University Press.

FRANCIS, G. & S. HUNSTON (1996) *Grammar Patterns 1: Verbs*. London: HarperCollins, COBUILD.

_____ (1998) *Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins, COBUILD.

FRANCIS, W. N. & H. KUCERA (1982) *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

GRANGER, S. (Org.) (1998) *Learner English on Computer*. New York: Longman.

HALLIDAY, M. A. K. (1991) Corpus studies and probabilistic grammar. In: K. AIJMER & B. ALTENBERG (org.). *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman.

_____. (1992) Language as system and language as instance: The corpus as a theoretical construct. In: J. SVARTVIK (org.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin, New York: De Gruyter.

HASAN, R. (1992) Rationality in everyday talk: From process to system. In: J. SVARTVIK (org.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin, New York: De Gruyter.

HENNOSTE, T. ET AL. (1998) Structure and usage of the Tartu University Corpus of Written Estonian. *International Journal of Corpus Linguistics*, 3.2: 279-304.

HOEY, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

_____. (1997) From concordance to text structure: New uses for computer corpora. In: B. LEWANDOSWKA-TOMASZCZYK & P. J. MELIA (org.). *PALC'97 – Practical Applications in Language Corpora*. Lodz: Lodz University Press.

HUNSTON, S. & G. FRANCIS (2000) *Pattern Grammar – A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/ Philadelphia: John Benjamins.

KENNEDY, G. (1991) 'Between' and 'through': The company they keep and the functions they serve. In: K. AIJMER & B. ALTENBERG (org.). *English Corpus Linguistics – Studies in honour of Jan Svartvik*. London / New York: Longman.

_____ (1998) *An introduction to Corpus Linguistics*. New York: Longman.

KJELLMER, G. (1994) *A dictionary of English collocations – Based on the Brown Corpus*. Oxford: Oxford University Press. (Three volumes)

LANGACKER, R. W. (1987) *Foundations of Cognitive Grammar – Volume I – Descriptive Applications*. Stanford, CA: Stanford University Press.

_____. (2000) A dynamic usage-based model. In: M. BARLOW & S. KEMMER (org.). *Usage-Based Models of Language*. Stanford: CSLI.

LEECH, G. (1966) *English in advertising: a linguistic study of advertising in Great Britain*. London: Longman.

_____ (1974) *Semantics*. Penguins: Harmondsworth.

_____. (1991) The state of the art in corpus linguistics. In: K. AIJMER & B. ALTENBERG (org.). *English corpus linguistics – Studies in honour of Jan Svartvik*. London: Longman.

_____. (1992) Corpora and theories of linguistic performance. In: J. SVARTVIK (org.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin, New York: De Gruyter.

_____. (1997) Introducing corpus annotation. In: R. GARSIDE et al (org.). *Corpus Annotation – Linguistic Information from Computer Text Corpora*. London and New York: Longman.

_____. (1999) Review of Biber, Conrad, and Reppen (1997) Corpus Linguistics – Investigating Language Structure and Use. *International Journal of Corpus Linguistics*, 4.1: 185-188.

MCENERY, T. & A. WILSON (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MOON, R. (1998) *Fixed Expressions and Idioms in English – A Corpus-Based Approach*. Oxford: Clarendon Press.

NATTINGER, J. R. & J. S. DECARRICO (1992) *Lexical phrases and language teaching*. Oxford: Oup.

OWEN, C. (1992) Corpus-based grammar and the Heineken effect: Lexico-grammatical description for language learners. *Applied Linguistics*, 14: 167-187.

PARTINGTON, A. (1998) *Patterns and Meanings – Using Corpora for English Language Research and Teaching* (Studies in Corpus Linguistics 2). Amsterdam/Philadelphia: John Benjamins.

PAWLEY, A. & H. SYDER (1983) Two puzzles for linguistic theory: Native-like selection and native-like fluency. In: J. RICHARDS & R. SCHMIDT (org.). *Language and Communication*. London: Longman.

PERCY, C. E. ET AL (Orgs.) (1996) *Synchronic Corpus Linguistics – Papers from the sixteenth International Conference on English Language and Research on Computerized Corpora (ICAME 16)*. Amsterdam/ Atlanta,GA: Rodipi.

PIKE, K. L. (1972) Towards a theory of the structure of human behavior. In: R. M. BREND (org.). *Kenneth L Pike – Selected writings*. Hague: Mouton.

QUIRK, R. ET AL (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

SANCHEZ, A. (1995) Definicion e historia de los corpus. In: A. SANCHEZ et al (org.). *CUMBRE – Corpus Linguistico de Espanol Contemporaneo*. Madrid: SGEL.

SANCHEZ, A. & P. CANTOS (1997a) El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus linguisticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y espanola y en cinco autores de ambas lenguas. *Atlantis*, 19.2: 1-27.

_____. (1997b) Predictability of word forms (types) and lemmas in linguistic corpora. A case study based on the analysis of the CUMBRE corpus: An 8-million word corpus of contemporary Spanish. *International Journal of Corpus Linguistics*, 2.2: 258-280.

SANCHEZ, A. ET AL (Orgs.) (1995) *CUMBRE – Corpus Linguistico del Espanol Contemporaneo – Fundamentos, Metodologia, y Aplicaciones*. Madrid: SGEL.

SCHÖNEFELD, D. (1999) Corpus Linguistics and cognitivism. *International Journal of Corpus Linguistics*, 4.1: 137-172.

SCOTT, M. (1997) PC Analysis of key words – and key key words. *System*, 25: 233-245.

SINCLAIR, J. McH. (1966) Beginning the study of lexis. In: C. E. BAZELL (org.). *In Memory of J R Firth*. London: Longman.

_____. (1987) Collocation: a progress report. In: R. STEELE & T. THREADGOLD (org.). *Language topics – Essays in honour of Michael Halliday* (Vol. 2). Amsterdam/Philadelphia: John Benjamins.

SINCLAIR, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

_____. (1995) From theory to practice. In: G. LEECH et al (org.). *Spoken English on Computer – Transcription, mark-up and application*. London: Longman.

_____. (1996) EAGLES Preliminary recommendations on Corpus Typology. EAGLES Document EAG TCWG CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale. Unpublished manuscript. Available at ftp:// ftp.ilc.pi.cnr.it.

SINCLAIR, J. McH. & A. RENOUF (1988) A lexical syllabus for language learning. In: R. CARTER & M. MCCARTHY (org.). *Vocabulary and language teaching*. London: Longman.

SINCLAIR, J. McH. ET AL (1987) *COBUILD English Dictionary*. London and Birmingham: Collins COBUILD.

STUBBS, M. (1995) Corpus evidence for norms of lexical collocation. In: G. COOK & B. SEIDLHOFER (org.). *Principle and Practice in Applied Linguistics – Studies in Honour of H Widdowson*. Oxford University Press: Oxford.

SVARTVIK, J. (Org.) (1992) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 – Stockholm, 4-8 August 1991* (Trends in Linguistics – Studies and Monographs: 65). Berlin, New York: Mouton De Gruyter.

THORNDIKE, E. L. (1921) *Teacher's Wordbook*. New York: Columbia Teachers College.

WEST, M. (1953) *A General Service List of English Words*. London: Longman.

YANG, D.-H. & M. SONG (1998) How much training data is required to remove data sparseness in statistical language learning? NLP Lab., Department of Computer Science, Yonsei University, Seoul, Korea, http:// december.yonsei.ac.kr/ ~dhyang.

ZHOU, Q. & S. YU (1997) Annotating the Contemporary Chinese Corpus. *International Journal of Corpus Linguistics*, 2.2: 199-238.