A LINGÜÍSTICA DE *CORPUS:* HISTÓRIA, PROBLEMAS, LEGITIMIDADE*

Jacqueline Léon**

RESUMO: A oferta de grandes *corpora* e a possibilidade de tratamento de enormes volumes de dados lingüísticos foram a origem, nos anos 1990, de uma renovação de interesse pelos tratamentos estatísticos e probabilísticos que, mais ou menos diretamente, questionam a lingüística sobre seu objeto, seus métodos e seus fundamentos. Esse interesse adquiriu uma importância crescente e tornou-se atualmente, sob o nome de "corpus linguistic", um campo de pesquisa dominante nas ciências da linguagem. Neste artigo, mostraremos que a denominação "corpus linguistics" recobre atualmente posições teóricas, temáticas de pesquisa e desenvolvimentos concretos muito heterogêneos. Examinaremos como a "*corpus linguistics*" inicialmente de origem britânica foi dotada posteriormente de uma legitimidade ao mesmo tempo histórica e teórica, pretendendo instituir-se como um novo paradigma nas ciências da linguagem. Distinguiremos duas atitudes no seio da tradição britânica: uma pretend*e erigi*r os estudos sobre corpus em novo paradigma e se apóia em uma construção retrospectiva das críticas de Chomsky dos anos 1950-60 para legitimá-la: a outra se inscreve na continuidade da tradição da lingüíst*ica empírica britâni*ca.

PALAVRAS-CHAVE: *Corpus linguistics*; Grandes *corpora*; História; Epistemologia; Chomsky; Lingüística britânica.

---

* Uma primeira versão deste texto foi objeto de uma conferência na USP em 25 de maio de 2005, a convite dos Programas de Pós-graduação em Filologia e Língua Portuguesa, Lingüística e Língua e Literatura Francesa. Agradeço às professoras Marli Quadros Leite, Esmeralda Negrão e Maria Sabina Kundman pelo convite. Essa conferência inscreve-se no quadro de um projeto de pesquisa da Fapesp, dirigido pela professora Marli Quadros Leite.

** Laboratoire d'histoire des théories linguistiques, CNRS, Université Paris 7, França

A oferta de grandes *corpora* e a possibilidade de tratamento de enormes volumes de dados lingüísticos foram a origem, nos anos 1990, de uma renovação do interesse pelos tratamentos estatísticos e probabilísticos que, mais ou menos diretamente, questionam a Lingüística sobre seu objeto, seus métodos e seus fundamentos. Esse interesse adquiriu uma importância crescente e tornou-se atualmente, sob o nome de "*Corpus Linguistics",* um campo de pesquisa dominante nas ciências da linguagem. Atestam esse fato a grande massa de anúncios regulares de obras ou de seminários e congressos nas listas de especialidades difundidas na Web, como a '*linguist list*' americana e as listas francesas RISC ou TALN.

Observe-se que a denominação "*Corpus Linguistics*", adotada hoje de modo unânime, é oriunda da corrente britânica, uma das mais antigas e estruturadas nesse campo. Grande produtora de manuais e de obras coletivas é a única a ser dotada de uma revista, a *International Journal of Corpus Linguistics,* que existe desde 1997, e uma coleção, *Studies in Corpus Linguistics* (publicada por Benjamins). Segundo Leech (1992), a primeira ocorrência do termo remonta à publicação de uma obra coletiva de 1984. No título, nota-se que o termo "*Corpus Linguistics*", ainda não banalizado, é acompanhado de um subtítulo explicativo: *Corpus Linguistics: recent developments in the use of Computer corpora in English Language Research.*

Esse termo pouco a pouco suplantou as outras denominações que mais designavam, no começo dos anos 1990, um conjunto de métodos do que correntes distintas, seja no domínio da descrição lingüística, do estudo da variação ou dos gêneros, ou, ainda, no domínio do tratamento automático das línguas*: corpus-based approach, corpus-based research into language, computational linguistics using large corpora, computer text corpora, statistical methods and large corpora.*[1]

[1] Os franceses, provavelmente divididos entre a tradição inglesa e a sua própria tradição de estudos estatísticos do vocabulário, falam de forma prudente de "lingüísticas de *corpus*" (no plural), ou de lingüísticas "sobre" *corpus*, o que aponta o fato de que se trata

A denominação "*Corpus Linguistics*" recobre atualmente posições teóricas, temáticas de pesquisa e desenvolvimentos concretos muito heterogêneos. Os trabalhos que usam a noção de *corpus* se distinguem (1) por concepções muito diferentes da noção mesma de *corpus*, (2) pelos objetivos (3) pelos domínios das Ciências da Linguagem aos quais eles remetem tanto quanto (4) pelos métodos de tratamento.

Na primeira parte deste artigo faremos um inventário dos diferentes aspectos dessa heterogeneidade, que tornam contestável a utilização de um termo único para designar trabalhos bem diversos.[2] Na segunda parte, examinaremos como a "*Corpus Linguistics*" de origem britânica foi dotada, posteriormente, de uma legitimidade ao mesmo tempo histórica e teórica, pretendendo instituir-se como um novo paradigma nas ciências da linguagem. Essa legitimidade baseia-se em uma postura ante a gramática gerativa. Efetivamente, é necessário distinguirem-se duas atitudes no seio da tradição britânica: uma pretende erigir os estudos sobre *corpus* em novo paradigma e se apóia sobre uma construção retrospectiva dos estudos sobre *corpus* e das críticas de Chomsky dos anos 1950-60, para legitimá-la; a outra se inscreve, sobretudo, na continuidade da tradição da lingüística empírica britânica sem necessidade de apelar para uma reconstrução histórica nem reivindicar uma postura teórica em ruptura radical.

1.1. Existem *corpora* de estatutos muito diferentes

A exaustividade definitória inicial, que vale, por exemplo, para os *corpora* de textos oriundos de línguas mortas, transformou-se para as línguas vivas em exigência de representatividade, quer se trate

antes de mais nada de método (Habert e outros 1997, Bilger 2000). Uma revista criada em 2002 se chama muito simplesmente "*Corpus*".

[2] A primeira parte deste artigo é a retomada parcial de um texto elaborado em colaboração com Marcel Cori e Sophie David no âmbito da preparação de um número da revista *Langages* consagrada à "Construção dos fatos na lingüística: o lugar do *corpus*", cuja publicação está prevista para junho de 2007.

de dados escritos, quer orais. Essa representatividade evoluiu também em razão dos avanços metodológicos e dos progressos técnicos: as coletas "por amostragem" dos anos 1970 deram lugar a "*corpus* de referência". Paralelamente, vimos desenvolverem-se *corpora* anotados, ou seja, enriquecidos de informações de natureza morfológica, sintática, semântica, prosódica, etc.

Por outro lado, atitudes radicalmente opostas são observadas, no que diz respeito à construção de *corpus*. A utilização da Web como um terreno de coleta de dados não pressupõe nenhum procedimento de construção do *corpus*. Na outra extremidade, para os historiadores que trabalham em Análise do Discurso, a construção do *corpus* exige que se investiguem as relações entre *corpus* fechado e *corpus* aberto, entre *corpus* e *não-corpus*, notadamente pelas noções de contexto e de interdiscurso.

1.2. Dados ou métodos?

Podemos dizer que "*corpus*" é um termo ambíguo que se refere, ao mesmo tempo, a um conjunto de dados e a um conjunto de métodos. No primeiro sentido (conjunto de dados), podemos dizer que todo lingüista é potencialmente um utilizador de *corpus*, pois a lingüística permanece de modo indiscutível, uma disciplina empírica. No segundo sentido (um conjunto de métodos), as pesquisas sobre *corpus* implicam métodos indutivos mais que hipotético-dedutivos, o que quer dizer que as análises conduzidas pelos dados (data-driven) são preferidas àquelas conduzidas por regras (rule-driven). Além disso, esses métodos são freqüentemente estatísticos e/ou probabilísticos, embora se possam encontrar também métodos sobre *corpus* fundados a partir de simples concordância.

1.3. Os objetivos

Os trabalhos que utilizam *corpus* podem ter objetivos muito diferentes:

• **Colocar os *corpora* à disposição**

Uma primeira tarefa pode consistir em colocar à disposição *corpus* para a comunidade de lingüistas, o que pressupõe uma reflexão sobre as escolhas de delimitação, transcrição, anotação. Os primeiros *corpora*, datados dos anos 1960 e reivindicados como pioneiros pela tradição "*Corpus Linguistics",* são essencialmente anglófonos, tais como o SEU (Survey of English Usage da University College of London), o Brown Corpus da Universidade Brown, dos Estados Unidos, mesmo se são às vezes elaborados em colaboração com Universidades não anglófonas – norueguesa, sueca ou holandesa – como o *corpus* LOB (Lancaster-Oslo-Bergen) e o *corpus* LUND (London-Lund). Todavia, na mesma época, um centro francês, o Centre du Trésor de la Langue Française (TLF), é criado em 1960 com o objetivo de elaborar um dicionário da língua francesa, a partir de um *corpus* que reunia todos os textos da literatura francesa de 1789 aos nossos dias (CNRS, 1961).[3]

• **Confeccionar instrumentos lingüísticos**

Um segundo objetivo é confeccionar instrumentos lingüísticos elaborados com a ajuda de *corpus*: bases de dados, dicionários, gramáticas, etc. Podemos citar em particular os dicionários e gramáticas elaborados a partir do *corpus* COBUILD de Birmingham (Sinclair, 1987).

• **Efetuar descrições lingüísticas**

Um certo número de lingüistas utiliza os *corpora* para efetuar descrições lingüísticas. Várias atitudes são possíveis: ao se adotar um procedimento indutivo, o objetivo consiste em descobrir fatos

[3] É preciso observar que, na maioria das vezes, o Brown Corpus é considerado pioneiro. Halliday (2002) é um dos raros a atribuir anterioridade ao SEU. O TLF não é mencionado por Kennedy (1998), em seu capítulo 2, consagrado à história dos *corpora* eletrônicos, senão como projeto conjunto franco-americano, que data dos anos 1980. Sobre essa questão, consultar Léon (2005a).

lingüísticos *nos corpora*, ou em descrever formas a partir de seu uso (prática denominada *corpus-based grammatical investigations*). Numerosos trabalhos consistem em fornecer informações quantitativas sobre a distribuição dos traços lingüísticos do inglês, segundo os gêneros (consultar os trabalhos reunidos ou descritos em Aijmer e Altenberg 1991, Johansson e Stenstrom 1991, Svartvik 1992, Kennedy 1998); ao se conservar um procedimento hipotético-dedutivo, o *corpus* é utilizado para compensar a falta de intuição ou completar dados estudados cada vez que a competência do lingüista se encontra deficitária. É o caso do estudo da entonação ou da morfologia derivacional, em particular o estudo da produtividade dos afixos (Fradin et al., 2003).

- **Estudos da variação**

A recorrência a um *corpus* permite criarem-se novos meios para trabalhar algumas questões ligadas à variação: os gêneros ou ainda a mudança lingüística (Biber, 1995).

- **Instrumentos de tratamento automático das línguas**

Enfim, é possível construir instrumentos de tratamento automático das línguas à base de *corpus*, em particular de *corpora* de controle que servem para parametrizar algoritmos probabilísticos, como o reconhecimento da fala, a correção ortográfica, a etiquetagem automática ou instrumentos de auxílio à tradução.

1.4. Domínios relacionados das ciências da linguagem

Os domínios das ciências da linguagem relacionados com os *corpora* são eles mesmos diversos: a lexicologia e a lexicografia, em particular o estudo das colocações; o ensino de línguas fundados sobre o uso; a terminologia; a sociolingüística: os trabalhos relativos à variação, aos gêneros e à mudança lingüística; a lingüística histórica; a informação e as sublinguagens. No domínio da Análise

do Discurso, o *corpus* é utilizado por especialistas de outras disciplinas a partir de hipóteses externas à Lingüística.

Não obstante essas abordagens muito diferentes, os atores do domínio, cada vez mais numerosos, tentam fornecer bases teóricas unificadas ao que se apresenta como uma crítica de fundo do programa de pesquisa da gramática gerativa, procurando, de certa maneira, fundar uma "nova lingüística". Inúmeros argumentos têm por finalidade distinguir os modelos probabilísticos dos modelos formais, seja porque os autores optem por uma oposição clara, seja porque defendam uma complementaridade ou até, uma "reconciliação" entre os dois enfoques. A posição que preconiza reconciliação tem como origem os trabalhos de Harris (1988), por uma Lingüística que associa teoria da informação e gramática formal (Abney, 1996, Manning, 2002; Pereira, 2000). Outros, como Biber, Conrad e Reppen (1998) preconizam uma complementaridade entre o enfoque intuitivo e o enfoque por *corpus*. Essa é também a posição de Kennedy (1998), que podemos considerar como próxima da corrente britânica.

Neste artigo, interessamo-nos, mais particularmente, por essa corrente, na medida em que ao mesmo tempo pioneira e dominante ela ilustra, de modo exemplar, os objetivos e as ambigüidades atuais da "*Corpus Linguistics*". Com efeito, podemos distinguir duas posturas. Falaremos de postura ou de atitude, já que, longe de formar duas correntes verdadeiramente distintas, elas colaboram com as mesmas publicações, e também porque os protagonistas, todos mais ou menos oriundos da London School, foram, diretamente ou indiretamente, discípulos de Daniel Jones e J.R. Firth. A primeira postura tem por ambição promover a "*Corpus Linguistics"* como novo paradigma para as ciências da linguagem. Ela é representada particularmente por Geoffrey Leech, um dos chefes da área, discípulo de Quirk – ele mesmo aluno de Firth –, diretor do *corpus* LOB, e, atualmente, professor emérito da Universidade de Lancaster. A segunda se inscreve mais na continuidade da lingüística empírica, cara à tradição britânica. Ela é representada em particular por MAK

Halliday, que foi aluno de Firth, John McH Sinclair, que trabalhou muito com Halliday na Universidade de Edinburgh, antes de constituir seu próprio grupo de pesquisa na Universidade de Birmingham, ou Graeme Kennedy, atualmente professor da Universidade de Wellington (Nova-Zelândia).

2. *CORPUS LINGUISTICS*: UM NOVO PARADIGMA PARA AS CIÊNCIAS DA LINGUAGEM?

2.1. Reconstrução e legitimidade histórica

Para Kennedy (1998), os avanços não são tão grandes como pretendem alguns, e a "*Corpus linguistics*" não constitui, certamente, um novo paradigma:

> Although there have been spectacular advances in the development and use of electronic *corpora*, the essential nature of text-based linguistic studies has not necessarily changed as much as is sometimes suggested. Corpus linguistics did not begin with the development of computers but there is no doubt that computers have given corpus linguistics a huge boost by reducing much of the drudgery of text-based linguistic description and vastly increasing the size of the databases used for analysis (Kennedy, 1998, p. 2).

Para os autores da primeira postura, ao contrário, a "*Corpus Linguistics*" é um enfoque filosófico novo, ou até, um conceito maior para as ciências da linguagem, como testemunham as citações de Leech (1992) e Stubbs (1997):

> **(1)**
> I wish to argue that computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject (Leech, 1992, p. 106-107).

> **(2)**
> First, corpus linguistics is a view about data: many different methods can be used to analyse corpus data. Second, a corpus is not just a tool, but a major concept in linguistic theory (Stubbs, 1997, p. 300).

Nos anos 1990, os partidários da "nova lingüística" propõem uma história das pesquisas sobre *corpus* que pode ser resumida do seguinte modo: nos anos 1940-50, a análise de *corpus* era próspera entre os estruturalistas americanos. Nos anos 1950, por suas críticas, Chomsky interrompeu essas pesquisas, que só puderam ser retomadas nos anos 1980-90, graças ao desenvolvimento inédito dos computadores, o que permitiu aumentar a memória disponível e a acessibilidade aos dados. Essa história, largamente admitida, é particularmente acreditada por Leech:

> **(3)**
> The impact of Chomskyan linguistics was to place the methods associated with CCL [Computer Corpus Linguistics] in a backwater, where they were neglected for a quarter of a century (Leech, 1992, p. 110).

> **(4)**
> The discontinuity can be located fairly precisely in the late 1950s. Chomsky had effectively put to flight the corpus linguistics of the earlier generation (Leech, 1991, p. 8).

Essa história se situa de modo mais geral dentro de um movimento que reivindica, nos anos 1990, no âmago do tratamento automático das línguas (Natural Language Processing), o ressurgimento do empirismo contra o racionalismo.[4] Os estudos de *corpus* são então reivindicados como uma redescoberta dos métodos empíricos e estatísticos dos anos 1950, no momento da difusão da teoria da informação de Shannon para o conjunto das disciplinas científicas. É inicialmente no domínio do reconhecimento da fala que, nos anos 1980, esses métodos foram de novo aplicados ao estudo da linguagem antes de serem utilizados em outras áreas da Lingüística, quando os métodos conduzidos por regras ou por conhecimentos (*knowledge-based and rule-based methods*) não mais satisfizeram.

Uma versão um pouco diferente é proposta pelo próprio Leech em 1991, quando ele evoca o surgimento, nos anos 1960, de uma

[4] Podemos nos referir à introdução do número especial da revista *Computational Linguistics* consagrada a "Computational Linguistics Using Large *Corpora*" (Church e Mercer, 1993).

segunda geração de *corpora*, intermediária entre os *corpora* dos neobloomfieldianos dos anos 1940-50 e aqueles dos anos 1990: o Survey of English Usage (SEU) de Randolph Quirk (Quirk, 1960) e o *corpus* Brown de Kucera e Francis (Kucera e Francis, 1967).

As conseqüências dessa construção histórica são importantes. Chomsky é apresentado como único responsável pelo desaparecimento dos *corpora* durante 30 anos. Reivindicando uma continuidade entre os trabalhos neobloomfieldianos, os *corpora* dos anos 1960, em particular o Brown Corpus, erigido pioneiramente, e os trabalhos atuais, Leech deixa entender que as críticas de Chomsky incidiram tanto no Corpus Brown quanto nos métodos neobloomfieldianos, e isto pelas mesmas razões. Enfim, os argumentos de Chomsky contra os *corpora* e os métodos estatísticos são invocados nos anos 1990 para instituir a "*Corpus Linguistics*" ao mesmo tempo como renovação dos estudos empíricos de *corpus* dos anos 1940-50 e como novo paradigma das ciências da linguagem.

A posição de Kennedy é, sob esse ponto de vista, mais moderada:

> The Brown Corpus was significant not only because it was the first computer corpus compiled for linguistic research, but also because it was compiled in the face of massive indifference if not outright hostility from those who espoused the conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky … Chomsky, among others, argued against the use of *corpora* and statistically based, probabilistic models of competence derived from linguistic performance. (Kennedy, 1998, p. 23)

Os estudos de *corpus* não estão certamente em primeiro plano nos anos 1960, mas não são considerados por Chomsky como o inimigo número 1. Uma vez colocados de lado os neobloomfieldianos, ele não manifesta, com relação aos estudos estatísticos em geral, senão alguns sarcasmos ou, na pior das hipóteses, uma certa indiferença. Além disso, ele não era o único na época a criticar os métodos empíricos. Isto é o que mostra um exame cuidadoso dos argumentos de Chomsky.

2.2. Os argumentos de Chomsky contra os *corpora*, as estatísticas e as probabilidades

2.2.1. Os *corpora*

Os *corpora* estão no âmago do enfoque neobloomfieldiano, para o qual o objetivo da Lingüística é uma taxinomia sistemática dos elementos lingüísticos em classes distribucionais, pela utilização de métodos empíricos, ou seja, a utilização de procedimentos indutivos de descoberta, a partir de um *corpus* de dados observados. Para Chomsky, esses procedimentos não revelam outra coisa senão fenômenos de superfície ao fornecer apenas um inventário estatístico de signos, desprovido de todo interesse explicativo.

É preciso, todavia, especificar que Chomsky não nega a necessidade de utilizar um *corpus* de dados observados, em particular por ocasião do estudo de línguas não descritas, como as línguas ameríndias. Mas esse *corpus* deve ocupar um lugar particular na maquinaria gerativa. Em um debate que o opõe aos neobloomfieldianos, por ocasião de um seminário organizado por Archibald Hill, na Universidade do Arizona em 1958, Chomsky expõe sua concepção hipotético-dedutiva de *corpus*:

> **(5)**
> Suppose I am working with an informant in a language which I do not know. I have gotten from the informant responses that tell me some formulations or guesses were good, some were not good. I also have in mind, from some source, a general theory of linguistic structure. This tells me what is the general form of grammars. I will revise my general theory whenever it turns out that there is a better formulation. As the result of a lot of operating with the data which I have now collected, I come out with a theory, a grammar of the proper form, which fits in with my general conception of grammatical forms. My grammar tells me that some things should be sentences, some should not. I go back to my informant, and try them out. If the informant agrees with my predictions, then I am content. How I got the theory in the first place is something I don't know. This is not properly a question belonging to the field of linguistics, it seems to me (Chomsky, 1962, p. 175).

Para estudar uma língua que não conhece, diz ele, o lingüista deve partir de um *corpus* natural de frases fornecidas por um informante nativo. Em uma segunda etapa, um outro *corpus* é gerado

pela gramática. Esse *corpus* vai conter ao mesmo tempo frases bem formadas e frases mal formadas. Ele será submetido ao teste do informante, o que vai permitir validar a gramática e a teoria. O *corpus* aparece, então, em duas etapas da máquina hipotético-dedutiva chomskiana: como entrada (*input*), que deve ser analisada pela teoria, e como resultado (*output*), gerado pela gramática. Apenas o *corpus* de entrada é "natural".

O objetivo da Lingüística é, no entanto, o conjunto de frases geradas pela gramática. Esse conjunto – o *corpus* – projetado pela gramática é infinito, contrariamente à concepção dos neobloomfieldianos, segundo a qual a língua é um conjunto finito de enunciados (utterances). Essa posição aparece em um dos primeiros artigos de Chomsky, "Three models for the description of language", publicado em 1956:

> **(6)**
> Similarly, a grammar is based on a finite number of observed sentences (the linguist's corpus) and it "projects" this set to an infinite set of grammatical sentences by establishing general "laws" (grammatical rules) framed in such hypothetical constructs as the particular phonemes, words, phrases, and so on, of the language under analysis. A properly formulated grammar should determine unambiguously the set of grammatical sentences (Chomsky, 1956, p. 113).

É de Nelson Goodman que Chomsky empresta a noção de projeção.[5] Em sua obra *Fact, Fiction and Forecast,* publicada em 1955, Nelson Goodman propõe uma teoria da projeção, segundo a qual as propriedades de uma amostragem podem ser projetadas por indução para o conjunto da população. A projeção é um método preditivo, e Chomsky sugere que o conjunto infinito das frases gramaticais seja projetado pela gramática a partir do *corpus* de dados observados na entrada.

[5] Sobre a noção de projeção herdada de Goodman, consultar Bourdeau (1970).

Chomsky trata também dessa questão no *Syntactic Structures*. A gramática projeta, a partir de um *corpus* de dados observados, o conjunto das frases gramaticais. Essa concepção é ligada à criatividade lingüística, a saber, a capacidade de o locutor produzir ou compreender um número infinito de frases:

> **(7)**
> First, it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances obtained by the linguist in his field work. Any grammar of a language will *project* the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances. In this respect, a grammar mirrors the behavior of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences (Chomsky, 1957, p. 15).

Essa concepção não é compartilhada pelos neobloomfieldianos, como mais uma vez se pode constatar, por ocasião da discussão na Universidade do Arizona. Chomsky declara que todo *corpus* natural é enviesado (skewed): ele não pode ser gerado, pois pode compreender não-frases (*non-sentences* ou *ill-formed sentences*) e ser incompleto:

> **(8)**
> HILL: It seems to me that if I were working with transformations, I would first select a representative sample of English sentences for my corpus. I would then try to see if by selection of kernel sentences within the corpus I could then generate the whole of the corpus. This is all that I would do.
> CHOMSKY: It is almost impossible to generate a corpus without going beyond it. Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.
> HATCHER: I have a corpus of about one hundred and twenty-five thousand sentences, and I do not find that it is skewed.
> CHOMSKY: But you do not have a machine which generates all of your sentences. I don't believe you could get a machine which would generate just these sentences. If you want to generate just the corpus and nothing beyond it, it would be a miracle if you could give any description shorter than the corpus itself (Chomsky, 1962, p. 159-60).

2.2.2. Criatividade lingüística, memória e inatismo da linguagem

Vimos no extrato (7) de *Syntactic Structures* que a projeção do *corpus* é estreitamente ligada à criatividade lingüística. Mais adiante no texto, ele postula que a freqüência de uso não intervém na parte da criatividade lingüística consagrada ao reconhecimento de frases gramaticais:

> **(9)**
> In the context "I saw a fragile –," the words "whale" and "of" may have equal (i.e., zero) frequency in the past linguistic experience of a speaker who will immediately recognize that one of these substitutions, but not the other, gives a grammatical sentence (Chomsky, 1957, p. 16).

Essas idéias são retomadas um pouco mais tarde na resenha que Chomsky fez em 1959 da obra de B. F. Skinner *Verbal Behavior*, em que ele precisa que a faculdade de reconhecer as frases (bem formadas) não é de ordem formal, semântica ou estatística, mas da ordem da criação infinita. Não é por comparação com as frases que teríamos guardadas na memória que reconhecemos uma frase:

> **(10)**
> We constantly read and hear new sequences of words, recognize them as sentences, and understand them. It is easy to show that the new events that we accept and understand as sentences are not related to those with which we are familiar by any simple notion of formal (or semantic or statistical) similarity or identity of grammatical frame (Chomsky, 1959, p. 56).

Foi em 1964, quando da publicação das atas do 9º Congresso Internacional de Lingüistas, que a criatividade lingüística se torna central na teoria chomskiana, em estreita relação com a infinitude e o inatismo da linguagem. Esse tema é introduzido no começo do texto. Além da capacidade dos locutores de produzir e compreender frases inéditas, a criatividade lingüística consiste também em identificar frases mal formadas e dar-lhes uma interpretação.[6] Nesse ponto, ain-

[6] Joseph (2003) assinala a assimetria da noção chomskianna de criatividade lingüística, mais centrada na produção do locutor do que na compreensão do ouvinte. O ouvinte se

da, a rememoração automática (*rote recall*) e a aprendizagem por decoração não ocupam senão um lugar restrito no uso da língua:

> (11)
> The central fact to which any significant linguistic theory must address itself is this: a mature speaker can produce a new sentence of his language on the appropriate occasion, and other speakers can understand it immediately, though it is equally new to them. … Normal mastery of a language involves not only the ability to understand immediately an indefinite number of entirely new sentences, but also the ability to identify deviant sentences and, on occasion, to impose an interpretation on them. It is evident that rote recall is a factor of minute importance in ordinary use of language, that "a minimum of the sentences which we utter is learnt by heart as such – that most of them, on the contrary, are composed on the spur of the moment", and that "one of the fundamental errors of the old science of language was to deal with all human utterances, as long as they remain constant to the common usage, as with something merely reproduced by memory" (Paul, 1886, 97-8). A theory of language that neglects this "creative" aspect of language is of only marginal interest (Chomsky, [1962] 1964, p. 914-15).

Mais adiante, no mesmo texto, ele precisa que a criatividade lingüística implicada no uso da linguagem é regida por um sistema de regras gerativas 'rule-governed creativity' mais que por variação 'rule-changing creativity' (Chomsky, [1962] 1964, p. 921). Argumento repetido nos *Aspects of the Theory of Syntax*, em que ele nega o papel da rememoração no uso da língua:

> **(12)**
> the fundamental fact about the normal use of language, namely the speaker's ability to produce and understand instantly new sentences that are not similar to those heard in any physically defined sense, or in terms of any notion of frames or classes of elements, nor associated with those previously heard by conditioning, nor obtainable from them by any sort of "generalization" known to psychology or philosophy (Chomsky,1965, p. 57).

---

limita a registrar passivamente as produções do locutor. Dois mecanismos estão operando na compreensão, segundo se trate de frases bem ou mal formadas: as frases bem formadas são interpretadas direta e automaticamente, enquanto impomos freqüentemente uma interpretação às frases mal formadas.

2.2.3. Os argumentos de Chomsky contra as estatísticas e as probabilidades

Como a maior parte dos cientistas de sua época, Chomsky se interessou pela teoria da informação de Shannon assim como pelos trabalhos sobre os modelos estatísticos do vocabulário de Zipf e de Mandelbrot.[7] Mas, desde 1956, ele rejeita toda definição de gramaticalidade em termos da lei de Zipf:

> **(13)**
> There is no significant correlation between order of approximation and grammaticalness. If we order the strings of a given length in terms of order of approximation to English, we shall find both grammatical and ungrammatical strings scattered throughout the list… (Chomsky,1956, p. 116).

Encontramos o mesmo argumento em *Syntactic Structures*:

> **(14)**
> If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list: there appears to be no particular relation between order of approximation and grammaticalness (Chomsky, 1957, p. 17).

Além disso, Chomsky duvida que frases muito simples, mas pouco utilizadas, possam ser encontradas em um *corpus*. É absolutamente necessário distinguir probabilidade e gramaticalidade:[8]

> **(15)**
> CHOMSKY: … I think "John ate a sandwich" is a highly unusual sentence that I would be unlikely to say in a lifetime. Just as I would be unlikely to say

---

[7] A lei de Zipf está baseada numa constatação empírica: quando as palavras de um texto estão ordenadas segundo sua freqüência decrescente, a freqüência de uma palavra é inversamente proporcional a sua posição. Benoît Mandelbrot elaborou um modelo estatístico propondo uma explicação teórica dessa lei.

[8] Halliday relata com certa amargura os sarcasmos de Chomsky, que datam da mesma época, sobre a utilização dos *corpora* e das freqüências. Estes devem, a nosso ver, ser reinterpretados, como o exemplo acima (15), a saber, que a gramaticalidade não pode ter fundamento probabilístico: "Chomsky's theory of competence and performance had driven a massive wedge between the system and the instance, making it impossible by definition that

> "grass is green," or "birds fly.". These sentences have zero probability. Maybe in talking about probability of sentences you mean grammaticality.
> STOCKWELL: You might say "John is eating a sandwich," but not "John eats a sandwich."
> CHOMSKY: Probability has to do with the number of times you find a given item. If we take a sentence like "John ate a sandwich," I would bet that you would not find it in all the sentences recorded in the Library of Congress (Chomsky [1958], 1962, p. 180).
> É preciso notar que a posição de Chomsky variou sobre a utilização das probabilidades. Ele parecia mais favorável a esse tipo de método no momento em que trabalhava com George Miller. No artigo comum deles, de 1963, compartilha a idéia de que a lei de Zipf deve ser levada a sério e que os resultados merecem ser examinados:

> **(16)**
> Miller and Newman (1958) have verified the prediction that the average frequency of words of length *i* is a reciprocal function of their average rank with respect to increasing length (Miller e Chomsky, 1963, p. 461).

Em troca, em sua resenha da obra de Vitold Belevitch, intitulada *Langage des machines et langage humain*, ele parece nitidamente menos entusiasta e sua posição sobre os trabalhos de Mandelbrot é ambígua: ao mesmo tempo em que coloca em dúvida a contribuição real da lei de Zipf para a Lingüística, reconhece que Mandelbrot tirou dela o poder explicativo:

> **(17)**
> The real import of Mandelbrot's work for linguistics seems to be that it shows that rank-frequency distributions of the type that Zipf and others have found are consistent with a very wide class of plausible assumptions about linguistic structure, and consequently, that we learn practically nothing about words when we discover this rank-frequency relation. In other words, this way of looking at linguistic data is apparently not a very fruitful one (Chomsky, 1958, p. 102).

---

analysis of actual texts could play any part in explaining the grammar of a language – let alone in formulating a general linguistic theory. Explicitly rejected was the relevance of any kind of quantitative data. Chomsky's sarcastic observation that '*I live in New York* is more frequent than *I live in Dayton Ohio'* was designed to demolish the conception that relative frequency in text might have any theoretical significance. [Linguistic Society of America Summer Institute, 1964]" (Halliday, 1991, p. 30).

Ele conclui de modo também ambivalente, colocando em dúvida o poder explicativo dos estudos estatísticos para a Lingüística, ao mesmo tempo em que afirma que estes podem ser dignos de interesse:

> **(18)**
> Although statistical properties of language and explanatory models for observed uniformities are certainly worth studying, it seems that such investigations have not yet reached the point where they make a significant contribution to the understanding of linguistic processes (Chomsky, 1958, p. 105).

Na realidade, sua posição é a seguinte: ele não nega o interesse dos estudos estatísticos, quando eles não são aplicados à sintaxe. Uma coisa é certa: Chomsky não está diretamente interessado por esses estudos, como testemunham os extratos seguintes de *Syntactic Structures,* em que aparece a questão da autonomia da sintaxe em relação à semântica e à freqüência de uso:

> **(19)**
> Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances. I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure (Chomsky, 1957, p. 17).

> **(20)**
> Given the grammar of a language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the use of language (as distinct from the syntactic structure of language) can be quite rewarding (Chomsky, 1957, p. 17, note 4).

2.2.4. Os modelos de Markov

Chomsky criticou muitas vezes a utilização dos modelos de Markov na sintaxe. Segundo esse enfoque, retomado por Shannon para a linguagem, a frase é um processo em estado finito esquerda–

direita, em que a probabilidade de ocorrência de uma palavra é determinada pela probabilidade de ocorrência das palavras que a precedem. No *Syntactic Structures*, Chomsky considera que esse modelo é incapaz de gerar um conjunto de frases gramaticais, e que é arriscado realizar-se por meio dele frases mal-formadas.

> **(21)**
> In short, the approach to the analysis of grammaticalness suggested here in terms of a finite state Markov process that produces sentences from left to right, appears to lead to a dead end just as surely as the proposals rejected in §2. If a grammar of this type produces all English sentences, it will produce many non-sentences as well. If it produces only English sentences, we can be sure that there will be an infinite number of true sentences, false sentences, reasonable questions, etc., which it simply will not produce (Chomsky, 1957, p. 24).

No entanto, Chomsky não desconsidera que esse modelo preditivo possa ser utilizado para elementos de nível inferior, como as palavras ou as letras (Chomsky e Miller, 1963). Dito isto, como Chomsky não se interessava pela palavra, a questão não é central para ele:

> **(22)**
> Higher-order approximations to the statistical structure of English have been used to manipulate the apparent meaningfulness of letter and word sequences as a variable in psychological experiments. As *k* increases, the sequences of symbols take on a more familiar look and – although they remain nonsensical – the fact seems to be empirically established that they become easier to perceive and to remember correctly. …We know that the sequences produced by *k*-limited Markov sources cannot converge on the set of grammatical utterances as *k* increases because there are many grammatical sentences that are never uttered and so could not be represented in any estimation of transitional probabilities (Miller and Chomsky, 1963, p. 429).

O essencial dos argumentos de Chomsky contra a utilização dos *corpora* e das estatísticas em Lingüística poderia se resumir na seguinte fórmula: as estatísticas, sim, com a condição de que elas não sejam concernentes à sintaxe e à frase. Vimos que a questão dos

corpora e dos métodos estatísticos está estreitamente ligada a opções centrais da GGT: autonomia da sintaxe, inatismo e infinitude da linguagem, função mínima da memória, da freqüência e das unidades pré-fabricadas na criatividade lingüística.

3. OS ARGUMENTOS DE CHOMSKY REVISTOS PELOS "*CORPUS LINGUISTS*"

3.1. Os partidários da construção retrospectiva

Um primeiro problema surge se considerarmos, segundo o que sugere Leech, que os argumentos de Chomsky são relativos ao Corpus Brown. Há, inicialmente, uma incompatibilidade cronológica: os argumentos de Chomsky datam dos anos 1956-65, enquanto a obra de Francis e Kucera foi publicada em 1967.

Além disso, segundo depoimento dos próprios autores, a obra comportava exclusivamente contagens de vocabulário a partir de extratos de textos, catalogados segundo os gêneros.[9] Segundo Sinclair (1991. p. 23), o principal interesse desse *corpus* consiste em seu método de coleta que não se mostra, contudo, válido senão para as palavras de freqüência elevada em um grande número de gêneros, considerando-se o tamanho limitado (2.000 palavras) de cada amostra.[10] Nenhum traço, então, no Corpus Brown de taxinomia, de classes distribucionais nem de procedimentos de descoberta. É por isso que é difícil dizer que o Corpus Brown era uma continuação dos *corpora* estruturalistas americanos e que ele era atinente aos argumentos de Chomsky. Ao contrário, vimos que Chomsky não rejeitava os estudos estatísticos de vocabulário, mas, todavia, não atribuia a eles muito interesse – o que era essencialmente o Corpus Brown.

---

9 Ainda atualmente, Kennedy (1998, p. 159) reconhece que há ainda muito poucos trabalhos quantitativos sobre a estrutura sintática das frases efetuados a partir de corpora.

10 Outro de seu interesse pelo Brown Corpus consistiu, desde o início, em seu acesso gratuito aos pesquisadores.

É necessário acrescentar que um dos autores do corpus Brown, Kucera, em seu artigo de 1968, declara estar de acordo com Chomsky sobre a restrição da utilização dos modelos de Markov para unidades de baixo nível, letras ou palavras (Kucera e Monroe, 1968).

Além disso, Chomsky não era o único a criticar os procedimentos de descoberta e os métodos empíricos dos anos 1950-60. Por exemplo, Bar-Hillel (1960), que promove, desde 1954, sua própria "sintaxe operacional", fundada sobre a gramática categorial, era inteiramente oposto a esses métodos.

Se retomarmos os argumentos de Leech para promover a *Corpus linguistics* como "new philosophical approach", percebemos que eles são mais práticos ou técnicos do que teóricos, mesmo se o autor anuncia um programa que se opõe ponto por ponto ao modelo chomskiano: o estudo da performance mais que da competência; descrição lingüística mais do que a busca dos universais, utilização de modelos quantitativos e não somente qualitativos; ponto de vista mais empírico do que racionalista.

O ressurgimento dos *corpora*, associado estreitamente ao desenvolvimento dos computadores na década de 1990, tem, segundo Leech, um impacto teórico real. "The new master is the computer" diz ele em seu artigo de 1992 (p. 105). O termo *"Computer Corpus Linguistics*", que ele promove, recobre a utilização de grandes *corpora* informatizados para o estudo em grande escala de uso no domínio do léxico e da sintaxe. Ora, esse termo refere-se diretamente a um domínio prático, do Tratamento automático das línguas (Natural Language Processing), e não à Lingüística teórica. Assim, no que diz respeito aos modelos de Markov, Leech opõe um argumento prático à crítica de Chomsky: as gramáticas de estados-finitos são as mais adaptadas e as mais performantes para efetuar a etiquetagem gramatical automática dos textos. Trata-se aqui mais de engenharia lingüística do que de lingüística teórica:

> **(23)**
> One thing in favour of probabilistic language processing systems is that they are eminently *robust*. They are fallible, but they work; they produce a more or less accurate result, even on unrestricted input data, in a way that outperforms most rule-driven language modelling systems (Leech, 1991, p. 18).

Nesse extrato, o argumento invocado relativo à robustez dos sistemas é de ordem técnica pertencente ao tratamento automático da língua que é relativo à confiabilidade de um programa e ao número reduzido de erros que ele possa vir a cometer. Argumentos esses difíceis de serem promovidos como opções teóricas para fundar um novo paradigma nas ciências da linguagem.

Essa posição não é inteiramente compartilhada, já que lingüistas, como Abney (1996), reconhece que os modelos de Markov não são verdadeiramente adaptados ao tratamento da língua e que o que Chomsky neles criticava não era o fato de serem estatísticos, mas, sobretudo, o de propor um modelo de estado finito da gramática:

> I think even the most hardcore statistical types are willing to admit that Markov models represent a brute force approach, and are not an adequate basis for psychological models of language processing. However, the inadequacy of Markov models is not that they are statistical, but that they are statistical versions of finite-state automata! Each of Chomsky's arguments turns on the fact that Markov models are finite-state, not on the fact that they are stochastic. None of his criticisms are applicable to stochastic models generally (Abney, 1996, p. 20).

3.2. A tradição britânica herdada de Firth

Agora, se examinarmos a segunda postura, que se inscreve mais diretamente na continuidade da tradição britânica originária da Escola de Londres e dos trabalhos de Firth, percebemos que ela não precisa recorrer a uma reconstrução histórica que invoque as rupturas para se legitimar. Na seqüência da abordagem sistêmica de Firth e do lugar que ele atribui ao estudo do sentido, sua posição,

relativa a Chomsky, se inscreve naturalmente na crítica firthiana do estruturalismo.

• **Gramaticalidade e aceitabilidade**

De fato, desde o fim dos anos 1950, os pesquisadores discutem o enfoque chomskiano, notadamente as noções de gramaticalidade, de aceitabilidade e sua relação com a freqüência da aparição de estruturas no *corpus*. Em seu trabalho de tese, Halliday estudou a freqüência de classes sintáticas, em um dialeto chinês, a fim de determinar se o não aparecimento ou o fraco aparecimento de uma estrutura no *corpus* era casual, se era uma prova de sua não gramaticalidade, ou ainda uma prova de sua raridade (Halliday, 1959, p. 58).

Outros, nos anos 1960-70, discutem a distinção gramaticalidade/aceitabilidade proposta por Chomsky e desenvolvem testes para estudar a relação entre freqüência e julgamento de aceitabilidade (Quirk e Svartvik, 1966, Greenbaum, 1976).

• **Hipótese probabilística e o continuum léxico-gramática**

A posição de Halliday sobre o *corpus* implica duas hipóteses principais que questionam seriamente a teoria chomskianna:

1) O sistema lingüístico é fundamentalmente probabilístico.

(24)
frequency in text is the instantiation of probability in the grammar (Halliday, 1991, p. 30: Halliday, 1992, p. 66).

2) Não há diferença entre o léxico e a gramática:

(25)
I have always seen lexicogrammar as a unified phenomenon, a single level of "wording" of which lexis is the "most delicate" resolution (Halliday, 1991, p. 31)

Essas opções opuseram muito cedo Halliday a Chomsky. No decorrer da discussão que se seguiu à intervenção de Chomsky, por ocasião do *9º Congresso de Lingüistas*, em 1962, Halliday reconheceu o interesse da noção de gramaticalidade (*grammaticalness*), com a condição de que ela seja expressa em termos de grau, e não em termos de exclusividade entre frases bem e mal formadas, e que ela seja completada pela idéia de lexicalidade (*lexicalness*) – (consultar Chomsky, 1964, p. 989).

O léxico e a gramática formam um *continuum*: em uma extremidade se encontra o léxico que tem a propriedade de ser aberto, enquanto a gramática, na outra extremidade, contém classes fechadas. Para Halliday, as duas hipóteses, propriedades estatísticas das línguas e complementaridade entre léxico e gramática, são estreitamente associadas. Se aceitarmos a noção de léxico-gramática, não tem sentido aceitar a idéia de freqüência relativa para o léxico e negar toda validade estatística às formas gramaticais. Assim, ele propõe que a lei de Zipf seja generalizada para a sintaxe.

A posição central do léxico e o *continuum* gramática-léxico serão compartilhados por Sinclair em suas pesquisas lexicográficas, notadamente na sua retomada, a partir dos *corpora*, da noção firthiana de colocação.[11]

Ele opõe dois princípios de interpretação do sentido em um texto: o *open principle* e o *idiom principle*. Segundo o *open principle*, para cada unidade do texto (palavra, sintagma, frase) o locutor dispõe de uma grande escolha de possibilidades lexicais, limitadas somente por restrições gramaticais. Os textos seriam tipos de seqüências vazias a serem preenchidas por léxico. Essa concepção, geralmente admitida em Lingüística, é compatível com a posição chomskianna: os itens lexicais, em número infinito, são fornecidos na entrada para a gramática, que gera, com a ajuda de regras, combinações lexicais infinitas.

[11] Para os primeiros trabalhos sobre a centralidade do léxico, consultar Halliday (1966) e Sinclair (1966). Relativamente às colocações, sobretudo à obra de Sinclair (1991) que aqui nos referimos.

Ao contrário, segundo o *idiom principle*, o locutor não dispõe muitas vezes senão de uma escolha limitada de sintagmas em parte pré-fabricados (*semi-preconstructed phrases*), quer seja de expressões estereotipadas, expressões em vias de fixação (como *of course*, assim como *anyway* ou *maybe*, deveriam no final formar uma única palavra), quer seja de colocações cuja adjacência não é na realidade necessária.

• **Colocações e criatividade lingüística**

O *idiom principle* e o que ele implica deu lugar a discussões sobre a noção chomskiana de criatividade lingüística. Antes de tudo é difícil negar, tendo em vista a importância dos segmentos pré-fabricados no uso da linguagem, o papel da memória na aprendizagem e na produção da língua.

Segundo Kennedy (1998), a idéia de colocação, por estar estreitamente associada à predictibilidade dos elementos e à existência, na linguagem, de elementos pré-fabricados, questiona profundamente a noção de criatividade lingüística tal como a concebe Chomsky. Para Kennedy, pode-se muito bem conceber a utilização de elementos parcialmente lexicalizados sem, no entanto, questionar o caráter inovador da linguagem:

> **(26)**
> In characterizing language as innovative, Chomsky was here referring specifically to the sequencing of words in sentences and was arguing against the Skinnerian conceptualization of the sentence simply as a left-to-right finite state Markov process or verbal chain in which the probability of a word's occurrence was determined by the occurrence of the words preceding it. The generative approach to language tends to downplay the use of prefabricated, ready-made sequences of words, although there is no reason why many sentences cannot be treated as partially lexicalized rather than purely syntactically generated (Kennedy, 1998, p. 109).

Esse argumento é também antecipado por Joseph (2003). A idéia mesma da criatividade lingüística infinita obriga Chomsky a rejeitar todo modelo "colocacional" da linguagem, enquanto para Sinclair e

seus discípulos, o fenômeno das colocações não implica a ausência de criatividade. Além disso, a existência de segmentos pré-fabricados contribui para refutar uma separação nítida – no cérebro – entre léxico e regras de gramática.

> **(27)**
> (…) those contemporary British (applied) linguistics traditions that stem from the teaching of Firth take seriously the evidence that our processes of speaking, writing and understanding do not proceed word by word, but in larger "pre-packaged" chunks. This observation has important implications for how we imagine language being "stored" in the brain. The standard imagery has long been of a grammar and a lexicon in our heads. The idea is that in one part of our brains is an inventory of "atomic" words, understood as sound-meaning correspondences, and in another part are rules for putting the words together (Joseph, 2003, p. 127-128).

• ***Corpus* / intuição**

Sinclair (1991) critica o recurso único à intuição, incapaz de dar conta de um certo número de fenômenos. A intuição, afirma ele, não diz nada sobre o uso. Ela não é de nenhuma utilidade quando se trata de estabelecer distinções relativas ao uso de um grupo de palavras mais do que de outro, ou seja, não há equivalente do julgamento de gramaticalidade para o léxico.

Os estudos estatísticos sobre os grandes textos argumentam também, segundo ele, contra a intuição. Em um texto grande, o sentido das palavras mais freqüentes não é o sentido dado como prioritário pela intuição. Parece, no que diz respeito às palavras mais freqüentes, que é preciso falar de uso mais que de sentido, na medida em que o uso provoca uma deslexicalização progressiva das palavras freqüentes, uma redução de sua contribuição distintiva para o sentido.

Alguns lingüistas, como Kennedy (1998), próximo de Sinclair, pronunciam-se a favor de um enfoque misto que associa intuição e *corpus*, e até se aproximam de Chomsky, em certos pontos. Kennedy reconhece que os *corpora* não podem dar conta completamente do funcionamento da linguagem. Eles não permitem diferenciar as es-

truturas possíveis das estruturas impossíveis. Contrariamente a muitos dos "*Corpus Linguists*", Kennedy parece não praticar o culto do atestado e reconhece que o fato de um elemento não aparecer em um *corpus,* mesmo muito grande, não significa que não exista. Inversamente, a aparição de uma estrutura em um *corpus* não estabelece automaticamente sua gramaticalidade.

> **(28)**
> The use of both introspection and corpus-based analysis can contribute to linguistic analysis and description. *Corpora* cannot tell us everything about how a language works. For example, they cannot be used as a basis for stating what structures or processes are not possible (…) The fact that an item or structure does not appear in even the largest corpus does not necessarily mean that it cannot occur, but could suggest the corpus might be inadequate or the item infrequent. Neither does the fact that a construction occurs in a corpus necessarily establish its grammaticality. (…) Whether utterances which involve phonetic or syntactic reductions such as *where you going?, wannanother one?* or *Good that you got here early* have to be accounted for grammatically will probably depend in the final analysis on frequency of occurrence and intuitive judgments as to what is "normal" (Kennedy, 1998, p. 271-272).

CONCLUSÃO

O acesso da "*Corpus Linguistics*" como domínio autônomo, e até como nova lingüística, encontra vários obstáculos. A unificação, sob um mesmo termo, do conjunto de domínios em que os *corpora* são utilizados, passa por uma postura epistemológica a qual é difícil subscrever: erigir um objeto prático (um *corpus*) e os métodos em lugar e espaço de um objeto teórico. Essa postura leva igualmente, forjando uma história mais ou menos credível, a encontrar modos de legitimação contestáveis. Argumentar com o fato de que as críticas de Chomsky teriam estancado os trabalhos sobre *corpus* nos anos 1960, significa apresentar os estudos estatísticos de vocabulário como os herdeiros da corrente neobloomfieldiana, o que eles, manifestamente, não eram.

As propostas avançadas como definidoras do novo paradigma pretendem se opor, ponto por ponto, ao programa chomskiano. Ora, observamos que elas não são em nada próprias da corrente "*Corpus Linguistics*"; elas são, com efeito, muito parecidas com as formuladas por certos funcionalistas para se distinguir do modelo gerativista e do estruturalismo em geral.[12] Os funcionalistas, quanto a eles, são mais favoráveis a uma postura continuísta do que a uma posição radical. Eles reivindicam uma continuidade entre racionalismo e empirismo, em particular a complementaridade entre dados originários do uso e aqueles fornecidos pela intuição, e a adoção eventual de métodos estatísticos; eles postulam uma versão atenuada do inatismo e um *continuum* entre o universalismo e o relativismo. Eles contestam a pertinência para a análise de uma distinção entre competência e desempenho, entre o conhecimento que o locutor tem da gramática e seu conhecimento do uso dessa gramática. A utilização de *corpora* e de métodos estatísticos não são, para os funcionalistas, senão um dos aspectos do enfoque funcionalista que permanece, antes de tudo, favorável à utilização em grande escala de dados empíricos. Como vimos, tal atitude pode, igualmente, ser adotada por alguns dos "*Corpus Linguists*" ao reivindicar para si sua filiação firthiana.

De fato, como antecipávamos no começo deste artigo, a utilização de *corpora* informatizados atravessa o conjunto das ciências da linguagem, qualquer que seja o paradigma adotado. Ela não constitui absolutamente um novo paradigma. Os *corpora* muito grandes constituem uma oferta de dados inéditos disponíveis e da qual cada lingüista, qualquer que sejam suas opções teóricas, só pode se regozijar. Adotaremos então, mais facilmente, o ponto de vista retros-

[12] Consultar o artigo de Noonan (1999), que situa o modelo funcionalista – trata-se aqui do que ele chama de "West Coast Functionalism" – relativamente ao modelo estruturalista e ao modelo formalista, que examina os diferentes traços que lhes são comuns, ou que lhes são opostos. Enquanto os funcionalistas vêem no paradigma gerativista uma das formas do modelo estruturalista, os lingüistas de *corpus,* ao pretenderem ligar-se às correntes empiristas neobloomfieldianas, ao contrário, as opõem.

pectivo expresso por Sinclair (1991) ou Halliday (2002), segundo o qual, graças aos avanços tecnológicos – os *corpora* informatizados, mas também a instrumentalização em fonética e os meios de gravação – a Lingüística dispõe, enfim, de dados dignos desse nome, enquanto antes ela se encontrava na mesma situação da Física do fim do século XV, privada de observações empíricas.

As tentativas de legitimação de um domínio prático não são verdadeiramente novas no domínio das ciências da linguagem, seja pela construção de uma história ou pela escolha de um nome unificador (ver Cori e Léon, 2002). Os desenvolvimentos políticos, financeiros ou industriais concorrem com os aspectos puramente científicos. Isso foi visto com a tradução automática e, de um modo mais geral, com o tratamento automático das línguas.[13]

BIBLIOGRAFIA

KARIN, Aijmer & ALTENBERG, Bengt (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London & New York: Longman.

BIBER, D. (1995). *Dimensions of Register Variation: a Cross-Linguistic Perspective*. Cambridge: University Press.

____; CONRAD, Susan; REPPEN, Randi. (1998). *Corpus Linguistics: investigating language structure and use*. Cambridge: CUP.

BILGER, M. (éd.) (2000). Linguistique sur Corpus, Etudes et Réflexions, *Cahiers de l'Université de Perpignan*, n. 31.

BOURDEAU, Michel (1979). *Chomsky et la critique des théories behavioristes du langage*, Doctorat de 3ème cycle, Université Paris 1.

CHOMSKY, Noam. (1956). Three models for the description of language. *IRE Transactions on Information Theory* IT-2, p. 113-124.

____. (1957). *Syntactic Structures*. The Hague, Mouton.

____. (1958). Review of Vitold Belevitch Langage des machines et langage humain 1956, *Language* 34-1: 99-105.

____. (1962). "Transformational Approach to Syntax". In: HILL, A. A. (Ed.). *Third Texas Conference on Problems of Linguistic Analysis in English May 9-12, 1958*. Studies in American English, The University of Texas Austin, Texas, p. 124-58.

[13] Este texto foi traduzido do francês por Maria Sabina Kundman e Marli Quadros Leite.

____. (1964). "The Logical Basis of Linguistic Theory". *Proceedings of the 9th International Congress of Linguists*, 1962, ed. by Horace Lunt, 914-978. The Hague: Mouton.

____. (1965). *Aspects of the Theory of Syntax*, Cambridge. MIT.

____. and Miller George. 1963. "Introduction to the formal analysis of natural languages ". *Handbook of Mathematical Psychology*, ed. by D. Luce, R. Bush and E. Galanter, v. II, New York: Wiley, p. 269-321.

CHURCH, Kenneth, Robert L. Mercer. (1993). "Introduction to the special Issue on Computational Linguistics Using Large *Corpora"*. *Computational Linguistics* 19, p. 1-24.

CNRS. (1961). *Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles*. 12-16 novembre 1957. Paris: Ed. du CNRS.

CORI, Marcel et LÉON, Jacqueline. (2002). "La constitution du TAL. Etude historique des dénominations et des concepts", *Traitement Automatique des Langues*, n. 43-3, p. 21-55.

FRADIN, B.; HATHOUT, N.; MEUNIER,F. 2003. "La suffixation en -ET et la question de la productivité. *Langue française*, 140.

GOODMAN, Nelson. 1955. *Fact, Fiction and Forecast*, Indianapolis and New York: The Bobbs-Merrill Company Inc.

HABERT, Benoit; NAZARENKO, Adeline; SALEM, André (1997). Les linguistiques de corpus. Paris: Armand Colin & Masson.

HALLIDAY, M.A.K. (1966). "Lexis as a Linguistic Level". *In memory of J.R. Firth*, C.E. Bazell, J.C. Catford, M.A.K Halliday, R.H. Robins, (Ed.) Longmans, p. 148-62.

____. (1991). "Corpus Studies and Probabilistic Grammar". In: AIJMER & ALTENBERG (Ed.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London: Longman: 30-43.

____. (1992). "Language as System and Language as Instance: the Corpus as a theoretical Construct" *Directions in Corpus Linguistics* ed. by J. Svartvik, 61-77. Berlin: Mouton de Gruyter.

____. (2002). *On Grammar*, Vol 1 in the Collected works of MAK Halliday, ed. by Jonathan Webster, London and New York: Continuum.

HARRIS Z.S. (1988). *Language and information*, New York: Columbia University Press

JOHANSSON, Stig; STENSTROM, A.-B. (1991). *English Computer Corpora: selected papers and research guide.* Berlin: Mouton de Gruyter

JOSEPH, John E. (2003). "Rethinking linguistic creativity". *Rethinking Linguistics*, Haylay Davis Talbot Taylor (Ed.). London and New York: Routledge Curzon

KENNEDY, Graeme. (1998). *An Introduction to Corpus Linguistics*. London and New York: Longman.

KUCERA, Henry & FRANCIS, W. Nelson (1967). *Computational Analysis of Present Day American English*. Providence: Brown University Press.

KUCERA, Henry & MONROE, George K. (1968). "A comparative quantitative phonology of Russian, Czech and German". New York: American Elsevier Publ.

LEECH, Geoffrey. (1991). "The state of the art in corpus linguistics". *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Aijmer & Altenberg (eds.). London & New York: Longman, p. 8-29.

____. (1992). *Corpora* and theories of linguistic performance In: Svartvik Jan (Ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*, Berlin, New York: Mouton de Gruyter, p. 105 -122.

LÉON, Jacqueline. (2005[a]). "Claimed and unclaimed sources of Corpus Linguistics". *The Henry Sweet Society Bulletin of History of Linguistic Ideas*, n. 44, p. 34-48.

____. (2005b). "Empiricism versus Rationalism revisited. Current Corpus Linguistics and Chomsky's arguments against *corpora*, statistics and information theory in the 1950-1960s.", XVI. Internationales Kolloquium des *Studienkreis der Geschichte der Sprachwissenschaft*, Nicosia (Cyprus), 11- 13.02. 2005.

QUIRK, Randolph and Jan Svartvik. (1966). *Investigating linguistic acceptability*. The Hague: Mouton.

SINCLAIR, John McH. (1966). *Beginning the study of lexis*. In: BAZELL, Catford, Halliday, Robins (eds.) Memory of J. R. Firth. London: Longman, p. 410-30.

____. (Ed.). (1987). *Collins COBUILD English Language Dictionary*, London and Glasgow: Collins.

____. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

STUBBS, Michael (1997). Review of T. McEnery & A. Wilson. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996, *International Journal of Corpus Linguistics,* v. 2-2, p. 296-300.

SVARTVIK, Jan (Ed.). (1992). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. Berlin, New York: Mouton de Gruyter.

ABSTRACT: During the nineties, the accessibility of large corpora and the possibility of manipulation of enormous quantities of linguistic data was the origin of a renewal of interest in statistical and probability evidences that served to directly question linguistics about its objectives, methods and foundations. This interest gained increasing importance and became important currently under the name of corpus linguistics, a field of dominant research in language science. In this article we will show that the designation corpus linguistics covers considerably heterogeneous theoretical positions and research, topics. We show how corpus linguistics, originally of british origin, was later endowed with historical and theoretical legitimacy while at the same time intending to establish itself as a new paradigm in language science. Finally we distinguish two attitudes inside the british tradition: one, intending to build the studies on a corpus and in a new paradigm based on a retrospective construction of the critical works of chomsky during the years 1959 and 1960, which was intended to legitimize the studies; the other attitude involves the continuity of the tradition of british empirical linguistics.

KEYWORDS: *Corpus Linguistics;* Great *Corpora*; History; Epistemology; Chomsky; British Linguistic.