Coleta, transcrição e análise de produções orais

Mirian Rose Brum de Paula
Universidade Federal de Santa Maria (UFSM)
Gema Sanz Espinar
Universidad Autónoma de Madrid (UAM)

GRPESQ/ CNPq Discurso, História, Gênero e Identidade
Laboratório Corpus: fontes de estudos da linguagem
Equipe Dynamiques des langues,
UMR 7114 MoDyCo CNRS et Université de Paris-X-Nanterre

Para ver uma cosa hay que comprenderla.
El sillón presupone el cuerpo humano, sus articulaciones y partes;
las tijeras, el acto de cortar.
Qué decir de una lámpara o de un vehículo?
El salvaje no puede percibir la biblia del misionero;
el pasajero no ve el mismo cordaje que los hombre de a bordo.
Si viéramos realmente el universo, tal vez lo enterderíamos.
Jorge Luis Borges.

Introdução

O emprego de coleções de textos, transcrições ou gravações nos trabalhos concernentes à linguagem não é recente. De fato, a criação de concordâncias[1] (*concondances*) é

anterior à aparição e à utilização generalizada do computador, do gravador ou da máquina de escrever. As primeiras concordâncias foram realizadas com a Bíblia: o objetivo era comparar as diversas versões desse texto a fim de constituir uma versão editorial normatizada (GARRIGUES: 1994). Esse árduo trabalho era, evidentemente, efetuado à mão. A partir de um olhar retrospectivo, observamos que essa tarefa manual foi deixada de lado recentemente. Vale salientar que não era possível realizá-la de outra maneira. Atualmente, as novas tecnologias tornaram esse trabalho artesanal completamente obsoleto, pois existem grandes bases de dados disponíveis, ao público em geral, em disquetes, discos rígidos e/ou cd-roms. Além disso, há possibilidade de se obter novos programas de computador capazes de realizar buscas de palavras ou seqüências de palavras (que levavam dias, meses ou anos através do método manual) em alguns minutos. Mas por que razão a manipulação automática do *corpus* com o qual desejamos trabalhar é importante?

Tentando responder essa pergunta, trataremos de três etapas que envolvem o trabalho do pesquisador. A questão principal, que engloba a anterior, é a de focar aspectos relacionados à coleta, à transcrição e à análise dos dados coletados. Afinal, porque digitalizá-los e que importância eles têm no desenvolvimento de diferentes trabalhos acadêmicos sobre a linguagem?

Acrescentamos ainda que não trataremos de bancos de dados constituídos a partir da língua escrita, embora possamos citar, em algum momento, esse tipo de *corpus*. Nesse artigo, colocaremos em evidência, a produção oral, os aspectos teóricos e, principalmente, metodológicos que a concernem.

1. Corpora[2] e língua oral

Quando abordamos a produção oral, mergulhamos no domínio da *performance*, da realização lingüística submetida aos imponderáveis de uma tarefa que se desenvolve em tempo real (*on-line*). Em pesquisas cujo objeto é a oralidade, o pesquisador necessita prever problemas metodológicos e teóricos adicionais. De fato, a fim de empreendê-las precisamos notadamente efetuar uma coleta de dados e uma transcrição das gravações efetuadas. Graças às novas tecnologias, essas tarefas tornaram-se menos trabalhosas e o oral conquistou um espaço importante e crível dentro dos estudos sobre a linguagem.

O estudo da oralidade permite o acesso a sistemas lingüísticos imersos no ambiente em que eles se originam, se transformam ou desaparecem. A fim de evidenciar e dar visibilidade a esse trabalho, destacamos que

a) a linguagem é adquirida pelo intermédio da língua articulada (nossa língua materna);
b) a língua escrita, após séculos de tradição essencialmente oral, emergiu a partir da língua oral e
c) a escrita é uma sofisticação da língua oral, ao mesmo tempo em que é uma maneira artificial destinada a fixá-la.

Na obra intitulada **Les linguistiques de corpus**, Habert, Nazarenko e Salem tratam de modo tímido a problemática que envolve a constituição e o tratamento de corpus orais. Os autores justificam-se da seguinte maneira:

> Os *corpus* orais transcritos ainda são raros: a transcrição propriamente dita e as escolhas e os custos que ela compreende freiam seu desenvolvimento, mesmo se ele parece mais acelerado nesses últimos anos. (...). Parece também que o oral impõe níveis descritivos e ferramentas teóricas parcialmente distantes

daqueles tradicionalmente utilizados para a escrita (2001: 13). (Tradução nosssa)

Essa constatação é freqüente nos estudos lingüísticos. Há um maior número de dados coletados a partir da escrita, embora os resultados dos trabalhos sobre a produção oral possam influenciar positivamente no desenvolvimento de ferramentas informáticas que permitam estocar grandes quantidades de dados (os *corpora* eletrônicos *anotados*) e realizar buscas automáticas através de sistemas potentes (os *concordances*). Como os *concordances* tradicionais, os *concordances eletrônicos* permitem encontrar, dentro de um *corpus* textual, todas as ocorrências de uma palavra inseridas em seus respectivos contextos. A diferença marcante entre esses dois *concordances* diz respeito à facilidade e ao acesso rápido aos dados. Os programas atuais permitem a realização, através de uma simples manipulação, de buscas de palavras ou grupos de palavras em alguns segundos. Eles permitem, dentre outras possibilidades, a busca de exemplos lexicográficos, o estudo dos contextos em que uma palavra é empregada, a análise de uma palavra ou de um campo semântico no interior da obra de um autor. Eis um exemplo oriundo do corpus **Mitterrand 1** que contém as intervenções radiofônicas e televisivas do ex-presidente francês François Mitterrand durante o seu primeiro mandato presidencial.

ue la france qui a acquis,	je le crois, la confiance et le respect
ères personnels, aussi, et	je le crois, qui se réfèrent à la moral
cer des propositions pour,	je le crois, saisir le monde entier du
rté des facilités qui ont,	je le crois, sauvé le secteur du textil
ation de la fin du siècle.	je le crois tout à fait, sans quoi je n
n souvient aussi – cela est	je le crois, tout à fait, venu de consi
de la république: je suis,	je le crois, très fidèle à ce que je su
jours, j'ai observé avec,	je le crois, une grande patience, pour
ants que cela contribuera,	je le crois, utilement au redresement
bre de plans, j'ai donné-	je le crois vraiment - plus d'expansion
racheter le portrait. moi,	je le dessine tous les jours, par des a
ite, je l'ai dit à alger,	je le dirai à amman en jourdanie où je
dans le monde. la france,	je le dirai simplement, a déjà apporté

Nesse fragmento, encontramos a forma *je* (=eu) inserida em contextos diferentes.

O acesso rápido às amostras que desejamos observar otimiza a pesquisa empreendida e caracteriza os *concordances eletrônicos* ou *automáticos*. Além dessa manipulação automática dos dados, esses *concordances* permitem encontrar palavras que iniciam por seqüências de letras (*te* ou *cerv,* por exemplo), palavras que terminam com prefixos específicos (*agem*, entre outros), formas flexionadas de um mesmo radical (*penso, pensava, pensamos*) ou, ainda, seqüências de palavras (*no entanto, tanto quanto, seja... seja*).

Corpus, segundo John Sinclair (1996: 4) é "uma coleção de dados linguageiros selecionados e organizados segundo critérios lingüísticos explícitos a fim de servir de amostra da linguagem". Nesse trabalho, *corpus* adquire uma dimensão suplementar relacionada ao fato dele estar ou não disponível eletronicamente. Assim, entendemos *corpus* como um conjunto de textos cuja origem é conhecida (data, autores, etc.) e que se encontra digitalizado.

Atualmente, como mencionamos *supra*, a maioria dos *corporas* eletrônicos é constituída de textos cuja origem é a língua escrita. Os *corpora* orais são raros porque é necessário passar pelos processos de coleta e de transcrição, o que torna mais lenta a constituição desses documentos.

2. Da língua à palavra articulada

Desde que Ferdinand de Saussure estabeleceu a *ciência da língua*, muitas abordagens deixaram de lado os dados orais a fim de desenvolver esse estudo científico. Dois elementos foram cruciais para que uma mudança ocorresse no sentido de introduzir o oral como fato observável no seio da pesquisa sobre a linguagem articulada:

a) O método de validação de hipóteses em lingüística, que abriu o caminho à heterogeneidade e aos estudos empíricos efetuados a partir de dados autênticos (próprios à oralidade).

Saussure foi o grande inspirador de uma ciência lingüística que se interessava principalmente pela escrita como fonte do que é *sistemático* na língua, comum a todos os sujeitos que a falam. Noam Chomsky privilegiou a validação de hipóteses a partir da intuição dos locutores nativos de cada língua a ser estudada. Os pesquisadores que trabalham com a *lingüística de corpus* (franc.: *linguistique de corpus*; ing.: *corpus linguistics*) assumem evidentemente que a validação de hipóteses deve ser realizada a partir de dados empíricos. Além disso, têm como objetivo a construção de *corpus* longos, representativos e *anotados* para tornar fácil e rápida a consulta de grandes quantidades de ocorrências do fenômeno lingüístico que desejam analisar ou que já estão estudando.

b) O reconhecimento de uma língua oral ao lado da escrita.

Pesquisadores de diferentes domínios recorreram aos *corpora* orais para desenvolver seus trabalhos: psicólogos (sobretudo os psicanalistas), psicolingüistas (principalmente os que realizavam estudos sobre o funcionamento do cérebro através do filtro da linguagem), neurolingüistas (cientistas que estudavam patologias da linguagem de origem neurológica), estudiosos em aquisição da linguagem e em aquisição de línguas estrangeiras (cujos domínios sempre enfocaram o uso comum e homogêneo da língua).

Mais próximo da lingüística, destacamos o desenvolvimento de estudos de línguas sem tradição escrita. Representativos, desse caso, são os trabalhos etnolingüísticos, realizados sobretudo no início do século XX por estudiosos americanos como, por exemplo, Franz Boas e Edward Sapir. Esses cientistas da linguagem, interessados pelas línguas indígenas nativas do continente americano, contribuíram positivamente à realização de estudos sobre a língua oral.

Além desses pesquisadores, que representam disciplinas diferentes, os lingüistas contam com *corpora* orais constituídos através do desenvolvimento de trabalhos em fonética e fonologia.

Dentro desse contexto, quais seriam as particularidades do texto oral? Tentando responder essa questão, comentaremos de modo mais detalhado, a seguir, as três etapas apontadas no título desse artigo. Destacamos, antecipadamente, que sua especificidade pode ser evidenciada através dos aspectos intonativo e morfo-fonológico (BLANCHE-BENVENISTE: 95), da variação em relação à norma ou do caráter pragmático das produções, pois na oralidade "há sempre a presença do outro", ou seja, a relação dialógica é potencialmente presente, visto que a comunicação interpessoal é sempre possível (ENCREVÉ: 96).

3. Etapas fundamentais da pesquisa dos fatos orais da linguagem

Na corrente denominada *lingüística de corpus*, encontramos muito mais do que uma simples escolha metodológica em vistas de uma melhor descrição da língua ou do

desenvolvimento de dicionários. Nela, identificamos, igualmente, importantes pressupostos teóricos que servem para definir o que é língua, delimitar o objeto da lingüística e melhor compreender suas relações com outros domínios. A língua enquanto objeto vivo, enquanto instrumento de comunicação inscrito na esfera social, através do qual o indivíduo constrói uma idéia de si mesmo e do outro ou através do qual as crianças aprendem a linguagem, diz respeito a diferentes domínios do conhecimento. Os *corpora* eletrônicos servem também para que o pesquisador teste suas hipóteses, para que possa confrontar modelos às realizações lingüísticas efetivas (HABERT, NAZARENKO e SALEM: 1997). Dentre os pesquisadores que se interessam pela linguagem articulada, destacamos lingüistas, etnolingüístas, sociólogos, especialistas da aquisição e da interação, psicólogos e historiadores. Essas visões pluridisciplinares acerca de um mesmo objeto contribuíram para que emergissem novos domínios (a psicolingüística, a etnolingüística, a sociolingüística... a *lingüística de corpus*) nos quais o pesquisador se confronta com a língua, o discurso e o texto.

Mencionamos que os *corpora* eletrônicos devem ser suficientemente longos, representativos e *anotados* a fim de que as hipóteses formuladas possam ser cientificamente validadas. O que isso significa? O tamanho e a representatividade dependem da qualidade das informações coletadas. Abordaremos esse problema na parte destinada à *coleta de dados*. O termo *anotação* remete a um *valor acrescentado* (LEECH, 1997) ou ao *enriquecimento dos dados* (HABERT, NAZARENKO e SALEM: 1997), ou melhor, implica o acréscimo de informações e enriquecimento do texto através da *anotação* de marcas morfológicas e sintáticas das palavras ou expressões constantes no *corpus* tratado. Trata-se de um "aporte de informações de natureza *interpretativa* aos dados brutos" (VÉRONIS, 2000: 2). Assim, os enunciados, as palavras, os segmentos devem ser *anotados* ou *etiquetados* a fim de permitir a pesquisa automática. Segundo Véronis, a anotação de um *corpus* oral inicia com a transcrição.

A constituição de um *corpus* começa antes da coleta, implica planificação e tempo. Após a realização da coleta dos dados, é preciso transcrevê-los e analisá-los com a ajuda de ferramentas mais ou menos eficazes. Embora existam atualmente produtos comerciais de qualidade destinados ao tratamento da linguagem e da fala, evidenciamos que a transcrição pode ser efetuada através de um tratamento de texto simples. É possível fazer análises sistemáticas conseqüentes com as grades existentes no *Word* ou com bases de dados do tipo *Access*. Essas ferramentas servem para auxiliar a realização de cálculos estatísticos. A utilização de um programa como o CHILDES (cf. *infra*) permite a realização de diversas manipulações: transcrever, calcular freqüências ou utilizar uma ferramenta de *concordances*, por exemplo. O CHILDES é evidentemente adaptado às pesquisas quantitativas.

As ferramentas existentes são destinadas à análise automática dos sistemas lingüísticos que ocupam um lugar de destaque no mercado mundial das línguas. Com elas, é possível tratar automaticamente o inglês, o francês, o espanhol... mas não podemos esquecer que a análise de dados deve ser, antes de tudo, qualitativa.

As manipulações efetuadas no *corpus* de modo eletrônico deveriam sempre ser verificadas pelo pesquisador. Caso contrário, perdemos completamente o contato com os dados que queremos analisar. O tratamento automático é difícil quando as pesquisas envolvem sistemas lingüísticos em construção: a língua da criança ou o sistema lingüístico de aprendizes de línguas estrangeiras. Logo, é melhor falar de *análises mediadas pelo computador* (como ocorre na tradução) do que *análises automáticas*.

Enfim, estamos frente à especificidade do oral em todos os momentos da pesquisa (durante a coleta, a transcrição e a análise), como veremos a seguir.

3.1 Durante a coleta quando

- procuramos informantes;
- gravamos e devemos tomar decisões que envolvem as seguintes questões:

a) O *corpus* será constituído de tarefas comunicativas ou de conversações livres?
É necessário fazer algumas perguntas ao informante ou descrever a tarefa que será efetuada por ele, a fim de coletar um *corpus* mais ou menos homogêneo, que possa servir para o estabelecimento de comparações.
b) A gravação ocorrerá durante uma situação comunicativa "programada" ou durante uma situação em que o informante não sabe que sua produção será gravada?
Esta questão concerne à ética da lingüística de campo.
c) Devemos intervir ou não durante a gravação?
O *paradoxo do observador* emerge quando o pesquisador concebe a pesquisa, realiza as gravações e analisa os dados coletados;

- encontramos eventuais problemas técnicos, tais como a qualidade medíocre das gravações, a estocagem dos documentos sonoros e a utilização (ou não) de mais de um gravador.

3.2 Durante a transcrição porque

- deve ser adaptada ao objetivo da pesquisa e a outros trabalhos que poderão ser desenvolvidos posteriormente. Normalmente, o *corpus* é coletado por pesquisadores que estão tratando um fenômeno preciso e as decisões concernentes à transcrição estão relacionadas a esse fenômeno. Atualmente, no entanto, os *corpora* começam a ser disponibilizados na rede para que possam ser reutilizados por outros lingüistas. É por essa razão que o pesquisador deve refletir acerca de determinadas escolhas, pois elas poderão restringir as análises que serão empreendidas a partir de um mesmo *corpus*. Uma transcrição ortográfica, por exemplo, não permitirá uma análise fonológica;
- existem três tipos de transcrição: fonética, fonológica e ortográfica (com ou sem relação com o documento sonoro, com ou sem relação com espectros acústicos);
- é necessário utilizar certas convenções destinadas à transcrição de fenômenos ligados ao caráter pragmático da situação conversacional, tais como entonação, auto-correções, pausas, trocas de turno, simultaneidade das falas, alongamentos de vogais, truncamentos bruscos, entre outros fenômenos ligados à comunicação interpessoal. Outras informações periféricas necessitam ser observadas: o papel dos interlocutores, as características do informante (idade, sexo, nome, profissão), os gestos e ruídos produzidos durante a interação.
A título de exemplo, destacamos duas maneiras de se realizar uma transcrição. A primeira foi adotada por Victorine Hancock (1997), da Universidade de Estocolmo, em um estudo sobre o emprego do conector macro-sintático *parce que*; a segunda, foi proposta por Claire Blanche-Benveniste (2000), num estudo sobre abordagens da língua francesa falada:

Victorine Hancock		**Claire Blanche-Benveniste**
E:;I:	Entrevistador; Informante	1. Elementos não ortográficos: apelo à notas e transcrições fonéticas
/ // ///	pausa curta, média e longa	2. Pontuação: nenhuma
+ SIM	marcas respectivas de início e de	3. Maiúsculas: somente nomes próprios,

	fim de enunciados que se sobrepõem	títulos de livros e filmes 4. Números: escrever por extenso
SIM	segue o discurso simultâneo do entrevistador	(exceção: números de telefones) 5. Pausas
(RISOS)	ruído não verbal	pausa curta: -
eh euh	hesitação	pausa longa: - -
X	sílaba incompreensível	interrupção: ///
:	sílaba alongada	6. Incompreensível: XXX (cada x corres-
NÃO	sílaba apoiada	ponde à discriminação de uma sílaba)
(I:mm)	sinal de retroação	7. Discursos simultâneos: .....
*	precede palavra transcodificada	.....
st	ruído efetuado com a língua	8. Multi-transcrição: /...../
$	fim de turno	9. Escolha ortográfica: (...)
		Ex.: nós somo(s) amigo(s)
		10. Retomadas: -
		Ex: muitos ca- casos sem solução

(Traduções nossas)

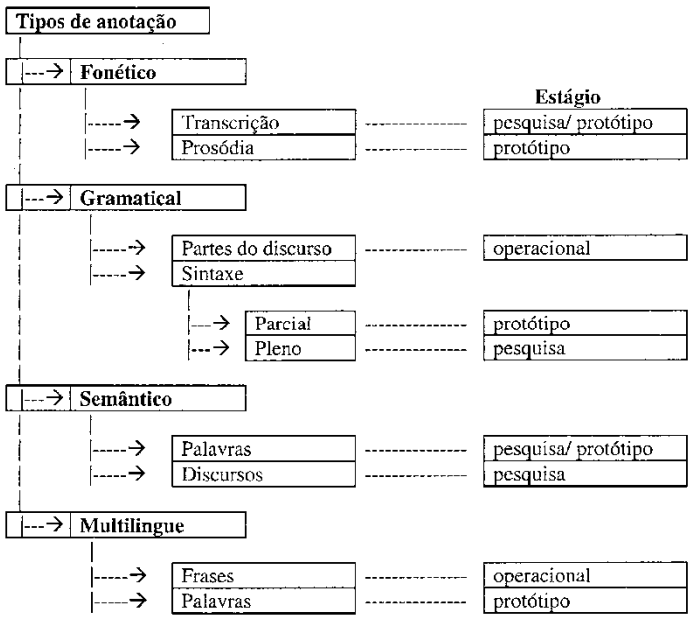
- não é possível submeter uma transcrição a um corretor ortográfico automático ou a um revisor, pois os dados coletados não são passíveis de correção. A *norma* e, principalmente, as normas da língua escrita não podem ser aplicadas ao domínio da oralidade;
- existem seqüências que não são transcritas devido à má qualidade da gravação sonora ou às ambigüidades oriundas do oral;
- as novas tecnologias facilitaram a realização de transcrições graças à possibilidade de estocar dados e de obter diferentes versões de um mesmo documento. Em decorrência disso, a mudança de critérios em relação à transcrição deve ocorrer antes de sua realização. Caso contrário, o novo tratamento que deverá ser aplicado ao *corpus* pode tornar esse trabalho lento e cansativo.
- é interessante nomear e acrescentar códigos para encontrar os documentos estocados. As produções transcritas devem ser segmentadas e numeradas. É por essas razões que consideramos a transcrição como primeira análise dos dados. De fato, a segmentação em enunciados é feita a partir de uma interpretação do discurso. A unidade enunciativa, no entanto, não é uma unidade de fácil identificação. Conseqüentemente, para alguns tipos de análise, pode ser interessante delimitar outras unidades: intonativas ou proposicionais, por exemplo, segundo o fenômeno focalizado.

3.3 Durante a análise, pois

- a anotação de traços orais apontará problemas oriundos da não naturalidade do discurso produzido, dos discursos não representativos ou dos *ratés* comunicativos;
- surgirão ambigüidades ou fragmentos de difícil interpretação mesmo quando estivermos escutando razoavelmente bem as palavras pronunciadas. A ambigüidade pode também ter sua origem na segmentação dos enunciados;
- não raro, novas palavras ou expressões são criadas pelo informante, há empréstimos lingüísticos, transferência de itens lexicais ou de expressões de uma língua para outra, no caso de produções em língua estrangeira;
- a *limpeza* do corpus talvez seja conveniente. Caso contrário, os cálculos globais poderão conter passagens desnecessárias. Por exemplo, se adotamos *xxx* para fazer referência às passagens não compreensíveis do *corpus*, isso será compreendido como uma palavra pelo programa a não ser que possamos prever uma espécie de *anti-*

dicionário contendo os itens lexicais que devem ser descartados no momento da realização da contagem quantitativa dos dados;
- a análise deve permitir o cálculo da freqüência de palavras, de expressões, de estruturas sintáticas ou semânticas;
- a análise deve ser empreendida em diversos níveis se queremos trabalhar o conjunto do texto (o que denominaremos perspectiva textual global). Dentre esses níveis destacamos o fonético, morfológico, sintático, semântico, pragmático, textual e enunciativo, embora possamos optar por um ou dois níveis de análise. Quanto à etiquetagem, ela pode ser efetuada de modo global ou específico. No caso de etiquetagem específica, anotamos o fato linguageiro que nos interessa.
Quanto a esse último tópico, é interessante acompanhar o estado da pesquisa concernente a diferentes tipos de anotações (etiquetagens) a fim de acompanhar sua evolução, conhecer e, talvez, adquirir, programas destinados à manipulação e à anotação automática de *corpora*.

No artigo intitulado "Anotação automática de corpus: panorama e estato da técnica" (2000), Véronis tentou dar conta dos diferentes tipos de anotação, verificando o desenvolvimento tecnológico de cada uma delas. O trabalho que empreendeu pode ser resumido, no quadro abaixo, proposto pelo pesquisador. Como podemos constatar, ele distingue três estágios relacionados à fase em que se encontra cada tipo de anotação: o material é *operacional* quando os programas já se encontram disponíveis no mercado. É um *protótipo* quando ainda está sendo testado. Nesse caso, ele é utilizado somente nos laboratórios de pesquisa. Enfim, é objeto de *pesquisa* quando existem trabalhos em andamento, mas não há emprego do material em situação de *anotação* real.



(Tradução nossa)

4. Disponibilidade e constituição de dados orais

Como vimos, os estudos sobre a linguagem a partir de bancos de dados existem há bastante tempo, mas a constituição de grandes arquivos somente pode ser realizada após o aparecimento e a comercialização de ferramentas básicas e atualmente muito comuns. Dentre elas, destacamos o gravador (cf. *supra*). Esse aparelho destinado ao armazenamento de informações e à reprodução sonora provocou a "caça ao documento autêntico" (CLAIRE-BENVENISTE e JEANJEAN: 43). Já a estocagem e a disponibilização de grandes arquivos de língua oral transcrita estão associadas às novas tecnologias: ao computador e aos programas destinados à manipulação e à anotação automática de dados textuais, principalmente.

As pesquisas sobre a linguagem efetuadas a partir de produções orais tendem a aumentar. Estudiosos têm criado grupos e redes interinstitucionais com o intuito de desenvolver importantes e ambiciosos projetos relativos à língua oral. Dentre eles, destacamos, no Brasil, o Projeto de Estudo da Norma Urbana Lingüística Culta (Projeto NURC) cujos pesquisadores coletaram dados de variantes cultas do português falado em São Paulo, Rio de Janeiro, Recife, Salvador e Porto Alegre e têm colocado à disposição da área de Letras publicações contendo elementos significativos visando à constituição de uma gramática referencial da variante culta do português do Brasil; na Europa e nos estados Unidos, colocamos em evidência quatro grandes projetos de pesquisa que contêm trabalhos empíricos efetuados a partir de dados orais: o projeto H.P.-D (*Heidelberger Forschungsprojekt "Pidgin Deutsch"*) dirigido, entre 1974 e 1986, por Klein e Dittmar sobre a aquisição do alemão por adultos, com pouca formação escolar e/ ou profissional, cujas línguas maternas eram o espanhol e o italiano; o projeto Z.I.S.A. (*Zweitspracherwerb Italienischer und Spanischer Arbeiter*), realizado entre 1975 e 1977, sob a coordenação de Meisel; as pesquisas sobre a aprendizagem do inglês por falantes de língua espanhola e por um adulto de origem cambojana realizadas pelos pesquisadores americanos Schumann (1978) e Huebner (1983), respectivamente; e o projeto E.S.F. (*Fondation Européenne de la Science*) sobre a aquisição de línguas estrangeiras (inglês, alemão, holandês, francês e sueco) por imigrantes falantes de diferentes línguas naturais (*pendjabi*, italiano, turco, árabe marroquino, espanhol e finlandês). O banco de dados produzido durante esse projeto é gerado pelo Max Planck Institut für Psycholinguistik de Nimèque. Acrescentamos, ainda, o *Child Language Exchange System* (plataforma CHILDES, já citado *supra*) desenvolvido, principalmente, por Macwhinney e Snow desde 1984. Esses últimos pesquisadores criaram um grande arquivo de dados, um sistema de transcrição e um conjunto de programas destinado à análise de produções naturalistas de sujeitos em fase de aprendizagem, ou seja, crianças bilíngües, indivíduos com patologias associadas à linguagem e aprendizes de línguas estrangeiras. Trata-se de um sistema computadorizado de intercâmbio de dados cuja função é a transcrição, codificação e análise do material lingüístico reunido. Além desses projetos, que contêm dados a partir de pesquisas empíricas e essencialmente orais, citamos o *corpus etiquetado* BNC (*British National Corpus*). O BNC compreende uma grande variedade de situações de comunicação que mistura produções orais (10%) e escritas. Trata-se do maior *corpus* oral do planeta!

Nem todos os dados recolhidos nesses projetos estão disponíveis. O sistema CHILDES contém resultados de aproximadamente "cem projetos de pesquisa sobre a linguagem em mais de uma dúzia de línguas, referentes aos últimos 25 anos" (MacWHINNEY e SNOW: 132). O material coletado durante o programa ESF também faz parte desse banco de dados.

> A disponibilidade desses dados confere, desde os anos 80, uma nova dinâmica às pesquisas formalistas (generativas) sobre a L2. Até então,

> essas pesquisas fundamentavam-se nos estudos transversais de aprendizes escolarizados (...) submetidos à tarefas experimentais que forneciam dados essencialmente institucionais. Os bancos de dados abrem a possibilidade de que outros pesquisadores adotem uma metodologia longitudinal. (PERDUE: 222-223). (Tradução nossa)

Como podemos observar, os diferentes bancos de dados orais selecionados para ilustrar esse trabalho não são constituídos de produções homogêneas: o tipo de texto (argumentativo, descritivo, narrativo), o perfil do informante (no que diz respeito ao sexo, à idade, ao grau de instrução, às motivações ou características do aprendiz, por exemplo), a língua transcrita (materna ou estrangeira) ou os fenômenos lingüísticos que são tratados podem ser completamente diferentes. O tipo de transcrição adotado pelo lingüista acompanha essa heterogeneidade, pois não existe uma maneira de se realizar uma transcrição, mas várias possibilidades de se transformar a língua oral em documento escrito.

É possível trabalhar a partir de um *corpus* já existente e coletado por outros pesquisadores (cf. *supra*). Para quem trabalha sobre a aquisição infantil ou sobre sistemas lingüísticos em desenvolvimento (ing. *learner variety*, franc. *lectes d'apprenants*, al.: *lernervarietät*), uma boa fonte de produções orais transcritas é o banco de dados CHILDES que pode ser recuperado via internet. Porém, o ideal para quem está aprendendo a pesquisar é passar pela fase da coleta de material lingüístico, pois o engajamento e a realização desse trabalho leva à reflexão e à tomada de decisões importantes para o desenvolvimento do projeto de pesquisa.

Em relação ao plano global da pesquisa, devemos definir o número de informantes, o tipo de locutor em função do que desejamos observar (uma ou várias línguas, textos *alinhados*[3], língua materna, língua estrangeira, língua padrão, variedade de uma região), da tarefa lingüística utilizada para suscitar a produção oral, a duração da gravação, o número de palavras (ou a extensão) do corpus que estamos constituindo.

Esse último item é importante porque pode corresponder a critérios de representatividade e servir na quantificação dos fatos da linguagem. Representar e quantificar são aspectos importantes no contexto da pesquisa[4] referente à língua oral.

Os trabalhos atuais realizados com algumas das línguas mais difundidas do planeta, como o inglês, o português, o espanhol e o francês, demonstram que a arquitetura dos bancos de dados textuais deve respeitar critérios de representatividade. O número de itens lexicais pode variar muito. A título de exemplo, destacamos a presença de:

- 100.000.000 palavras etiquetadas no **British National Corpus**;
- 2.000.000 no **Corpus Clef** do francês atual (Benoît Habert, CNRS);
- 1.767.163 no **Corpus de Referência do Português Contemporâneo** (CRPC);
- 1.100.000 palavras no *corpus* oral de referência do **Espanhol Contemporâneo Peninsular** (Marcos Marín, Universidade Autônoma de Madri);
- 435.000 palavras no corpus etiquetado **London-Lund**.
- 570.000 palavras em **Linguagem Falada** (Mark Davies, Illinos State University);
- 305.124 ocorrências no corpus etiquetado **Mitterrand 1** (D. Labbé, Institut d'Études Politiques de Grenoble);
- 273.070 palavras em **Arthus**, corpus misto de espanhol contemporâneo da Universidade de Santiago de Compostela (contém 18% de produções orais);

O tamanho do *corpus* depende dos objetivos da pesquisa, dos recursos humanos e meios econômicos disponíveis. Os objetivos podem ser muito diferentes. Se eles são abrangentes, é possível constituir um *corpus de referência*[5]. Caso os objetivos sejam mais

pontuais, a fim de que o *corpus* seja empregado para fins precisos possibilitando análises finas em fonética, lexicologia, análise da conversação, análise do discurso ou em aquisição, entre outras, é necessária a coleta de um *corpus especializado.*

O *corpus de referência* fornece informações profundas sobre o funcionamento de uma língua natural e pode representar todas as variedades pertinentes e todo o vocabulário característico dessa língua. Ele serve como suporte fundamental na elaboração de gramáticas e dicionários. O *corpus especializado* é limitado a uma situação comunicativa ou a um domínio específico.

Marcos Marín (1994) evidencia seis critérios concernentes à coleta, à transcrição e à anotação dos dados: a *oralidade*, a *espontaneidade*, a *adequação*, a *representatividade*, a *autenticidade* e o *standard.* Encontramos diversos problemas nos *corpora* orais relacionados a esses critérios de *cientificidade.* De acordo com o fato linguageiro pesquisado, é necessário, por exemplo, excluir os discursos cujo suporte é a língua escrita. Dentre eles, encontramos os discursos políticos, as comunicações, as conferências, as emissões de rádio ou televisivas, pois os locutores apoiam-se geralmente na escrita para elaborar seus textos orais. Além disso, é preciso encontrar o suporte adequado, aquele que possa servir à disponibilização e potencial reutilização dos dados coletados. Para tanto, o pesquisador deve visar tanto à padronização dos métodos relacionados às entrevistas, às transcrições e às etiquetagens quanto à explicitação dos critérios ligados à representatividade dos dados concernentes aos percentuais adequados para cada tipo de texto, por exemplo.

Enfim, não podemos confundir *língua oral* e *conversação.* O conceito de *conversação* faz alusão a um tipo de *gênero* discursivo e existem diferentes gêneros relacionados ao oral, como ocorre com a escrita. Marcos Marín propõe critérios para a distribuição dos tipos de discursos oral e escrito, em termos de percentuais para cada tipo, dentro de *corpus de referência.* Isso pode ser considerado como um *standard* no que concerne à representatividade quantitativa.

Para corpus orais, os percentuais seriam os seguintes: textos científicos (2-5), conversações (15-20), educativos (5-6), ciências humanas (5-10), parlamentares (4-6), jornalísticos (25-30), técnicos (10-15).

O protocolo da enquete deve se adaptar ao tipo de discurso que procuramos gravar. No caso da conversação, o entrevistador deve se engajar no papel indicado (deve ou não intervir, por exemplo). Quando desejamos estocar produções realizadas por crianças ou por aprendizes de línguas estrangeiras que tenham um nível ainda rudimentar, o entrevistador tem um papel imprescindível durante a coleta dos dados. É muito comum, nesses casos, que ele recorra à tarefas lingüísticas semi-controladas, ou melhor, a atividades que dão origem a textos muito próximos do monólogo. Nelas, o entrevistador procura não participar durante as gravações. Não há, conseqüentemente, muitas mudanças de turnos. Em gravações que privilegiam a interação face a face, a atitude do entrevistador é completamente diferente.

5. Transcrição do material

Em relação a outros tipos de produção, o texto oral é abundante, variável e, conseqüentemente, mais difícil de ser conservado, representado e manipulado. Quando realizamos uma transcrição, suprimimos informações ou acrescentamos elementos ao texto original. Duas dificuldades devem ser destacadas em relação a essa tarefa: dificuldades que têm sua origem na percepção, pois "escutar é uma atividade complexa [e] estamos sempre prontos a escutar o que acreditamos plausível" (BLANCHE-BENVENISTE e JEANJEAN: 6), e problemas relativos à legibilidade da transcrição, ou melhor, ao modo como ela será realizada a fim de que o pesquisador possa trabalhar confortavelmente e o leitor possa ter acesso rápido

aos dados. Por esse motivo, a transposição da produção oral para o papel merece atenção e cuidados especiais.

> O transcritor ingênuo será vítima de sua ignorância e de todos os fenômenos ligados à reconstrução; ouvinte não advertido, ele arrisca entender mal, mesmo tendo boa vontade (GOFFMAN, 1981: 214). É preciso lhe dar uma formação mínima [...]. Ele deve ter uma idéia referente ao objetivo da transcrição e deve poder centrar sua atenção nos aspectos que deseja particularmente estudar. Colecionar uma quantidade de dados e identificar somente depois o que será utilizado [e analisado] não é uma boa maneira de começar o trabalho (BLANCHE-BENVENISTE e JEANJEAN: 98) (Tradução nossa)

De fato, a fim de trabalhar com dados orais, é necessário selecionar um quadro teórico e metodológico que dê conta do fenômeno que desejamos analisar. O objetivo da pesquisa e a escolha do(s) aspecto(s) da linguagem que o pesquisador almeja estudar devem preceder a coleta de dados, pois o objetivo é de constituir um *corpus* que contenha uma alta freqüência dos fatos de linguagem selecionados. Caso contrário, o pesquisador corre o risco de perder o seu tempo e de engavetar as produções recolhidas. Em casos extremos, há inversão das primeiras etapas da pesquisa e o pesquisador, de posse do material gravado e, talvez, transcrito, reestrutura o seu trabalho em função do material que tem em mãos. Não raro, estudantes de iniciação científica e de pós-graduação não sabem o que fazer com os dados que reuniram. Esse tipo de problema ocorre em decorrência de uma certa negligência no que diz respeito aos aspectos metodológicos da pesquisa: quando "um pesquisador profissional ou iniciante tem grandes dificuldades no seu trabalho, isso ocorre quase sempre por razões de ordem metodológica" (QUIVY e VAN CAMPENHOUT: 4). Para evitar essa situação, é necessário conceber um método de trabalho e tentar respeitá-lo de modo regular e sistemático.

O sistema de transcrição escolhido (fonético, ortográfico, fonético acompanhado da versão ortográfica, intonativo), por exemplo, também deve fazer parte da metodologia adotada. Como a transformação de sons, ritmos, entonações, gestos e/ ou hesitações em escrita é uma atividade árdua, demanda paciência, homogeneidade no tratamento do material lingüístico e tempo daquele que a empreende, não deve ser efetuada de qualquer modo.

A transcrição não é uma operação mecânica, mas uma verdadeira reconstituição perceptiva das condições de produção, pois não é empreendida durante a situação comunicativa e a regulação intersubjetiva de seus participantes. Essa tarefa corresponde à primeira interpretação e simplificação dos dados recolhidos que passam pelo *filtro* da percepção do pesquisador e adquirem, aos poucos, características do texto escrito, mesmo quando o sistema de codificação ou transcrição utilizado consegue preservar, de modo mais ou menos fiel, as informações veiculadas pelo comportamento interacional do(s) informante(s). Alguns autores evidenciam essa transformação (do oral que se transforma em escrita) negando, inclusive, a possibilidade de se trabalhar a oralidade a partir desse tipo de documento.

> Durante muito tempo, os lingüistas trabalharam a oralidade após o término da transcrição [dos dados coletados]. [...] Se transcrevemos o oral, fazemos dele escrita. É preciso preservar toda a extraordinária especificidade do oral, todas as marcas que não encontram

> correspondentes na escrita, mesmo com o auxílio dos alfabetos fonéticos mais completos (ENCREVÉ: 104). (Tradução nossa)

Segundo Encrevé, o lingüista não percebe a diferença entre os textos oral e escrito porque, em geral, o acesso à oralidade realiza-se através do documento oral transcrito, ou seja, pelo intermédio da língua já representada. Esse tipo de argumento reforça alguns mitos e preconceitos em relação à legitimidade da língua falada enquanto objeto de pesquisa. Voltamos a enfatizar que tudo depende dos objetivos do lingüista e dos fenômenos linguageiros que ele deseja observar. Além disso, o tipo de transcrição adotado dependerá das intenções do pesquisador em relação à acessibilidade do material transcrito (ele deve levar em consideração o seu público alvo, se constituído de especialistas ou não) e à fidelidade ao material de origem (se ele tem ou não a intenção de respeitar a autenticidade da gravação).

A transcrição parece ser incontornável nos trabalhos que envolvem pesquisa de campo: há necessidade de se manipular facilmente os dados coletados. A freqüente manipulação dos dados engendra um sentimento de familiaridade com esse material, permitindo a emissão de hipóteses sobre o funcionamento da linguagem (restrito, evidentemente, ao *corpus* analisado). Mas que tipo de transcrição empreender? Baseada em alfabetos fonéticos ou no código escrito (ortográfico)? Contamos atualmente com ferramentas que permitem imbricar o som e sua transcrição. CHILDES é um bom exemplo desse tipo de ferramenta. De fato, não é difícil digitalizar o som e associar, na imagem que encontramos na tela do computador, o som, o espectro acústico e a *legenda*. Esse trabalho é útil nos estudos dos componentes fonológicos. Por outro lado, no que concerne aos trabalhos sobre a *gramática do oral*, o espectro sonoro não é necessário. Nesse caso, é suficiente realizar uma transcrição baseada na ortografia usual e acrescentar alguns signos complementares para marcar intonações ou pausas, por exemplo.

6. Análise de produções orais

Ao longo do século XX, em diferentes momentos e escolas lingüísticas, foram privilegiados os estudos da fonética (domínio inicialmente investigado pelos estruturalistas), da morfologia e da sintaxe (domínios trabalhados durante muito tempo pela gramática gerativa). Atualmente, em relação ao estudo dos fenômenos lingüísticos, parece claro que os *corpora* devem ser objetos polivalentes para servir de suporte de pesquisas que possam ser realizadas em diversos níveis: fonético, morfológico, sintático, lexical, semântico e/ou pragmático.

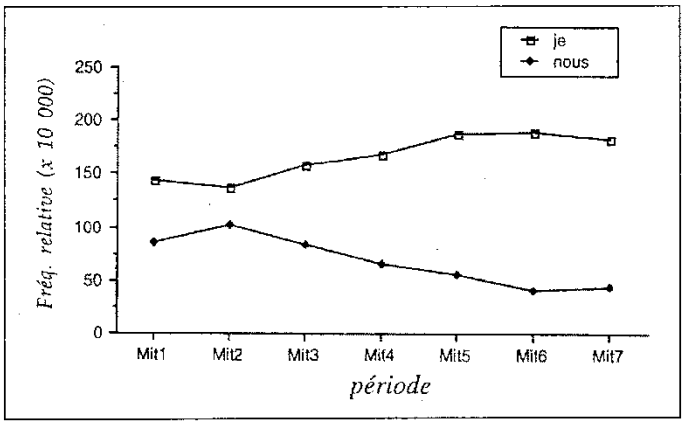
No que diz respeito aos *corpora de referência*, as anotações (ou etiquetas) devem responder às normas atuais de intercâmbio de documentos (formato SGML). Por exemplo, *<catgram>adv.</catgram>* poderia ser a etiqueta de *categoria gramatical*; *adv.*, seria a abreviação de advérbio. Essas etiquetas devem, de modo ideal, pertencer a níveis diferentes e possuir forma abreviada (três letras normalmente) correspondente ao campo da lingüística. As etiquetas poderiam compreender as informações seguintes:

a) incisos, interrogação (nível fonético),
b) *<catgram>*: adj., adv., v.; *<número>*: sing., pl.; *<gênero>* mas., fem. (nível morfológico),
c) sujeito, objeto (nível sintático relativo à função) e SNO, SN1, SN2 (nível sintático concernente à posição relativa no enunciado),
d) agente, paciente (nível semântico),
e) tópico, foco (nível enunciativo),
f) movimentos referenciais: introdução, manutenção, deslizamento (nível textual),
g) conversação, entrevista (nível tipológico).

Se todos esses códigos estivessem em formato SGML, teríamos mais facilidade na homogeinização das etiquetas quando analisamos diferentes *corpora*.

Destacamos anteriormente que há necessidade de se efetuar análises qualitativas e quantitativas. A corrente anti-empirista, anti-numérica e pro-simbólica dos últimos vinte anos descartou a quantificação dos dados. Segundo Liberman (1991), contar era considerado como atividade não apropriada "para uma pessoa de qualidade". Porque essa atividade é importante? Que fenômenos linguageiros podem ser melhor apreciados através da quantificação dos dados?

De uma maneira geral, sabemos que a constituição de listas exaustivas dos contextos em que o fenômeno pesquisado aparece faz surgir regularidades (cf. *concordances supra*) permitindo a generalização dos resultados da pesquisa. Com o auxílio de ferramentas automáticas, essas regularidades são identificadas rapidamente. Se esse processo fosse efetuado manualmente, não revelaria a sistematicidade e a regularidade de muitos desses fenômenos. Para que isso fique mais claro, citaremos um exemplo extraído de manipulações do corpus **Mitterrand 1**. O estudo em questão focaliza a repartição dos pronomes pessoais da primeira pessoa (*je* = eu e *nous* = nós) empregados pelo ex-presidente francês, em emissões de rádio e televisão, em cada um dos sete anos de seu primeiro mandato. Na figura abaixo, a primeira pessoa do singular está representada por ⊟ e a primeira pessoa do plural por ●):



Duas tendências podem ser evidenciadas a partir da quantificação efetuada. A primeira concerne aos seis primeiros anos de seu mandato: o *eu* aumenta e o *nós* diminui. A segunda, está relacionada com o último ano de seu primeiro governo: há inversão da primeira tendência, ou melhor, as ocorrências da primeira pessoa do singular começam a diminuir e as ocorrências da segunda pessoa do plural aumentam. Essas variações interessam especialistas do texto político. A perspectiva "quantitativa é aqui a única via de acesso à análise detalhada e contrastiva" (HABERT, NAZARENKO e SALEM: 186) desse tipo de fenômeno.

Estudos puramente qualitativos ou puramente quantitativos deveriam ser evitados, pois há complementariedade entre essas duas perspectivas.

Os estudos quantitativos podem ter um caráter estatístico forte. Os pesquisadores próximos a esses trabalhos utilizam, às vezes, medidas estastísticas elaboradas. Elas devem ser adaptadas às necessidades de cada domínio.

Nos trabalhos sociolingüísticos, por exemplo, há tratamento quantitativo do fenômeno relativo à variação. Uma das maneiras de abordar esse fenômeno consiste na criação de escalas implicacionais e no desenvolvimento de estudos estastísticos acerca da

distribuição de um certo fenômeno nas produções de um grupo de locutores. Esses estudos quantitativos devem ser completados por estudos qualitativos destinados a explicar a origem e os limites da variação.

Conclusão

Privilegiando a dimenção automática do tratamento de produções orais, tentamos abordar três macro-etapas do desenvolvimento de pesquisas cujo objeto é a língua falada: a coleta, a transcrição e a análise de dados. Essa primeira reflexão sobre essas etapas está relacionada com a importância metodológica do trabalho a ser empreendido. Ela versou sobre a necessidade de se distinguir tanto o tipo de dados e os fenômenos analisados como a maneira de os analisar, pois podemos utilizar dados orais sem respeitar uma abordagem pragmática da linguagem, ou colocando de lado questões ditas *tradicionais* como as que concernem à morfosintaxe. A nossa proposta visa integrar tudo isso, isto é, descrever e explicar fenômenos de cunho lingüístico. Isso implica trabalhar com unidades de natureza diferente: fonética, morfológica, lexical, sintática, enunciativa, textual e discursiva, entre outras. Logo, não se trata somente de coleta de dados e de transcrição bruta, mas de etiquetagem também, o que implica análises mediadas pelo computador. Através da ajuda do suporte eletrônico, é a análise completa dos textos que está em jogo. Essa importante e sedutora perspectiva objetiva motivar o pesquisador a lidar com a complexidade da linguagem articulada.

Notas

[1] Trabalho de identificação de todas as ocorrências de uma palavra dentro de um conjunto de dados textuais. Essas palavras, inseridas em seus contextos respectivos, são, em seguida, regrupadas. Trata-se, segundo o **Dicionário Hoaiss** (2001), de um "índice alfabético de vocábulos apresentados nos contextos em que aparecem (num trecho, num autor, numa época, etc.)" [oferecendo] "a possibilidade do estudo comparativo das palavras e dos diversos empregos do mesmo vocábulo".

[2] Referimo-nos normalmente a *corpus*, no singular, e *corpora*, no plural.

[3] O *corpus alinhado* é constituído de textos paralelos acompanhados de suas respectivas traduções.

[4] É possível verificar isso, entrando na rede web e digitando *corpus oral português*, *corpus oral español*, *corpus oral français*, *corpus oral english* a fim de encontrar informações sobres esses bancos de dados orais.

[5] Um corpus de referência é, segundo Marcos Marín (1991) uma grande base de dados textuais, ou melhor, diversas bases de dados interligadas, unidas em um sistema de estruturação de dados, de textos de referência e de ferramentas informáticas que servem para o tratamento dessas informações. A título de exemplo, citamos o projeto C-CORAL-ROM (Corpora de Referência Integrada para Línguas Romanas Orais) que procura disponibilizar cd-roms de quatro línguas romanas: espanhol, português, francês e italiano.

Referências Bibliográficas

BLANCHE-BENVENISTE, C. **Approches de la langue parlée en français**. Paris: Ophrys, 2000.

ENCREVÉ et al. Actualité de l'enquête et des études sur l'oral. **Langages** n° 93. Paris: Larousse, 1992.

BORGES, J. L. **El libro de arena**. Paris: Gallimard, 1990.
GARRIGUES, M. Concordances automatiques pour exercices authentiques. **Le Français dans Le Monde** n° 274. Paris: Hachette, 1994.
HABERT, B., NAZARENKO, A. e SALEM, A. **Les linguistiques de corpus**. Paris: Armand Colin, 1997.
HANCOCK, Victorine. Parce que: un connecteur macro-syntaxique. **Aile**. n° 9. Paris: Instaprint, 1997.
HOUAISS, A. e VILLAR, M. S. **Dicionário Houaiss da língua portuguesa**. Rio de Janeiro: Objetiva, 2001.
LEECH, G. Introduction corpus annotation. **Corpus annotation: Linguistic information from computer text corpora**. Londres: Longman, 1997.
MACWHINNEY, B. Análise computadorizada das interações. **Compêndio da linguagem da criança**. Porto Alegre: Artes Médicas, 1997.
MARCOS MARÍN, F. A. **Informática y humanidades**. Madri: Gredos, 1994.
PERDUE, C. E GAONAC'H, D. Acquisition des langues secondes. **L'acquisition du langage**, volume II. Paris: PUF, 2000.
QUIVY, R. e VAN CAMPENHOUDT, L. **Manuel de recherche em sciences sociales**. Paris: Dunod, 1988.
SINCLAIR, J. Preliminary, recommendations on Corpus Tpology. Relatório técnico EAGLES (Expert Advisory Group on Language Engineering standards), CEE, maio de 1996.
VÉRONIS, J. Annotation automatique de corpus: panorama et état de la technique. **Ingéniérie des langues**. Paris: Hermes Science Europe, 2000.