Parâmetros de compilação de um corpus oral: o caso do C-ORAL-BRASIL

Tommaso Raso (UFMG)
Heliana Mello (UFMG)

RESUMO: O artigo reporta os principais aspectos da compilação de um corpus de fala espontânea do português brasileiro, o C-ORAL-BRASIL, enfocando a tipologia textual que o compõe, os critérios de transcrição e segmentação da fala, os procedimentos metodológicos adotados, o treinamento de transcritores e os primeiros resultados descritivos já obtidos a partir de um estudo piloto.

Palavras-chave: Corpus oral; Português brasileiro; Enunciados; Unidades informacionais.

1. O projeto

Nesse trabalho apresentamos pela primeira vez[1] o *corpus* C-ORAL-BRASIL, *corpus* da fala espontânea do português do Brasil (PB), que está em construção dentro de um projeto coordenado por Tommaso Raso com a colaboração de Heliana Mello, em andamento na Universidade Federal de Minas Gerais. Ao longo deste artigo ficará claro que nem todas as fases do trabalho possuem o mesmo grau de definição, porém todas estão bem encaminhadas. O projeto é financiado pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela Universidade Federal de Minas Gerais (UFMG) e pelo Banco Santander[2]. O *corpus* já nasce com o objetivo de estudar a estrutura informacional do português do Brasil (PB) e suas ilocuções com base na Teoria da Língua em Ato (CRESTI 2000)[3], e constitui-se como a quinta

[1] Com a exceção de apresentações orais em congressos e seminários.
[2] Os financiamentos incluem capital para compra de equipamento e material bibliográfico, recursos para viagens da equipe do projeto e para visitas de pesquisadores externos, bolsas para pessoal técnico e alunos e pagamento de serviços de terceiros.
[3] Para o quadro teórico das ilocuções, vejam-se Austin, 1962; Cresti, 2005.

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 20-35 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

ramificação do C-ORAL-ROM (CRESTI e MONEGLIA, 2005), *corpora* de referência das quatro principais línguas românicas européias[4].

O projeto teve início, oficialmente, em janeiro de 2007, mas já antes, através de duas dissertações de mestrado[5], os coordenadores haviam elaborado um piloto do projeto, constando de dois textos orais (um dialógico informal e um monológico formal), com um total de 5395 palavras. No piloto estavam presentes todas as etapas metodológicas de compilação e tratamento de dados, i.e., gravação, transcrição, segmentação e etiquetagem informacional, além de haverem sido estudadas nos dois textos as principais medidas da fala e as unidades informacionais de tópico e apêndice[6]. Ao longo dos dois anos seguintes, outros objetivos foram acrescentados: o estudo da expressão da modalidade no PB, dentro do paradigma dos estudos do grupo LABLITA[7]; e a comparação entre o PB e o português europeu, com vistas, em particular, à identificação de aspectos do PB que possam enriquecer o arcabouço de estudos sobre evolução e contato lingüísticos. Hoje o projeto conta com um amplo grupo de colaboradores: três doutorandas, duas mestrandas, seis alunos de graduação, quatro dos quais bolsistas, e um estatístico. Vários trabalhos foram publicados[8], e a formação do corpus está em fase avançada. A previsão é de que até o final de 2009 a metade informal do *corpus* seja completada, sendo essa considerada a parte mais relevante para os estudos sobre a fala espontânea. Já a parte formal do *corpus* será iniciada em 2010.

2. O *corpus*: características gerais

O *corpus* prevê pelo menos trinta horas de gravações, das quais quinze de fala informal e quinze de fala formal. A metade formal ainda não foi completamente definida. Em princípio, será seguida a arquitetura do C-ORAL-ROM[9], mas é provável que algumas adaptações ao contexto sócio-lingüístico brasileiro sejam necessárias. As características da metade informal, já completamente definida e em fase avançada de realização, e seu processo de formação serão de fato o objeto deste trabalho.

A arquitetura da parte informal prevê um mínimo de quinze horas de gravação distribuídas em um mínimo de cem textos de, em média, 1500 palavras. Uma porcentagem reduzida de textos (não mais de 25) poderá ser constituída por textos maiores (4500 palavras) ou textos menores, desde que apresentem uma consistente autonomia textual. Dos pelo menos cem textos, 20% serão de contexto público e 80% de contexto familiar/particular. Em cada contexto, um terço dos textos será constituído de monólogos e dois terços de diálogos ou conversações (ou seja, diálogos com mais de dois participantes), buscando-se uma representatividade paritária entre essas duas últimas tipologias.

Como no C-ORAL-ROM, escolheu-se representar uma única diatopia. No nosso caso a diatopia do estado de Minas Gerais, e em particular a área urbana da capital, Belo Horizonte.

[4] Existe um acordo entre a Faculdade de Letras da UFMG e a Facoltà di Lettere dell'Università di Firenze coordenado por E. Cresti e T. Raso que prevê, entre outras coisas, a colaboração para a realização do *corpus* do português do Brasil e estudos interlingüísticos da estrutura informacional.
[5] Ulisses 2008 e Alves de Deus 2008.
[6] Os resultados foram publicados em Raso, Mello, de Deus e Jesus (2007), Raso e Ulisses (2008) e Raso e Mello (no prelo).
[7] Esse sub-projeto em particular é coordenado por H. Mello. Vejam-se Cresti 2002 e 2003.
[8] Além daqueles citados na nota 6, Maia Rocha, Raso e Andrade (2008).
[9] Cresti-Moneglia (2005); Cresti *et alii* (2004); Moneglia (2005).

Portanto, pelo menos 50% dos falantes incluídos no *corpus* serão mineiros, mas é muito provável que esse número no final resulte bem maior. Tradicionalmente a fala mineira é dividida em três grandes áreas[10]. Certamente todas serão representadas, mas não é objetivo do projeto a sistematicidade dessa representação; a grande área metropolitana da capital será certamente a de maior freqüência.

O *corpus* busca representar, em certa medida, a variação diastrática, que na fala brasileira é especialmente importante; mas também nesse caso, apesar de existir um esforço explícito para que várias faixas sócio-culturais estejam presentes, não se pretende alcançar uma representatividade estatisticamente confiável. Contudo, em todas as ramificações do *corpus* constarão interações com falantes das várias faixas sócio-culturais, tanto interagindo com falantes da mesma faixa quanto com falantes de faixas diferentes. Com relação aos critérios adotados para o C-ORAL-ROM, resolvemos adequar a indicação das faixas de escolaridade a uma partição que refletisse melhor a realidade brasileira, considerando as diferenças profundas entre ela e a realidade européia. Não é preciso dizer que há muitas diferenças significativas entre a sociedade brasileira e as sociedades européias. O quadro 1, a seguir, explicita as diferenças entre faixas adotadas no C-ORAL-ROM e no C-ORAL-BRASIL:

C-ORAL-ROM	C-ORAL-BRASIL
1 escola primária ou nenhuma escolarização	1(nenhum título de estudo: até o primário incompleto
2 segundo grau	2 até o título de terceiro grau, desde que não exerça uma profissão que necessite do título universitário
3 terceiro grau ou estudante universitário	3 profissão que necessita de título universitário ou superior
X desconhecido	X desconhecido

Quadro 1 – Faixas sócio-culturais

Essa distribuição objetiva refletir as diferenças entre o contexto sociolingüístico europeu e brasileiro quanto à relação entre nível escolar e nível lingüístico. Naturalmente nenhuma divisão tão esquemática pode ser plenamente satisfatória, mas acreditamos que nossa escolha de ampliar o espectro do estrato intermediário, com relação ao projeto europeu, represente melhor a realidade escolar brasileira. De fato, no C-ORAL-BRASIL a extensão do estrato intermediário (2) inclui uma parte dos estratos (1) e (3) do C-ORAL-ROM. Essa escolha visa a refletir a maior variabilidade entre as instituições educacionais brasileiras se comparadas àquelas européias.

A variação privilegiada e que se busca representar estatisticamente no *corpus* é a variação diafásica, por ser, na opinião dos coordenadores, aquela realmente significativa na variação estrutural da fala. Essa é a variação privilegiada na arquitetura em suas diversas ramificações: a divisão entre formal e informal; dentro do informal, a divisão entre contexto público e familiar/particular; dentro de cada contexto, a divisão em três tipologias interacionais

[10] Zágari (2005).

diferentes: monólogo, diálogo e conversação; dentro de cada tipologia interacional, a máxima variação possível de situações comunicativas.

Os textos são transcritos com o sistema CHILDES-CLAN, implementado pela notação prosódica com os critérios elaborados por Moneglia & Cresti (1997).

3. A variação diafásica

Cada tipologia interacional prevê o máximo de variação situacional possível. A monológica prevê uma variação tanto no assunto (contos de vida, entrevistas, monólogos de trabalho, piadas, narrativas, etc.), quanto nos destinatários dos monólogos (familiares, amigos, clientes, pessoas com as quais se mantém relações de trabalho, crianças, etc.), assim como nos lugares escolhidos para as gravações (lugar de trabalho, casa de amigos ou parentes, restaurante, etc.). Analogamente, nas conversações, e especialmente nos diálogos, além das diferentes relações entre os falantes e dos diferentes locais de gravação, buscou-se a maior variação possível quanto à atividade cumprida durante a interação: duas ou mais pessoas cozinhando, duas ou mais pessoas trabalhando juntas em um computador, uma pessoa explicando para outra(s) o funcionamento de um equipamento tecnológico ou de um programa de computador, cliente(s) interagindo com o vendedor em lojas, um pedreiro e um engenheiro tocando uma obra, duas pessoas fazendo compras no supermercado, dois garçons cuidando do serviço em uma festa, dois ou mais professores discutindo de trabalho, duas ou mais pessoas conversando em um carro, duas ou mais pessoas conversando amigavelmente em casa sem cumprir alguma atividade específica, duas ou mais pessoas em mesa de restaurante, duas ou mais pessoas fazendo contas, duas ou mais pessoas estudando juntas para uma prova, professor e aluno em uma aula particular, duas ou mais pessoas visitando um apartamento para ser alugado, etc. Assim, a variação situacional, dentro de cada ramificação, é dada pela combinação das seguintes variáveis, em ordem de importância:

- tipo de atividade cumprida durante a interação. Mudando-se a atividade, muda-se a situação comunicativa;

número e tipologia dos participantes da atividade. Mudando-se o número e/ou a tipologia dos participantes, muda-se a situação;

- lugar da interação. Mudando-se o lugar da interação, muda-se o tipo de situação;
- assunto da interação. A mudança de assunto contribui para a mudança situacional, mesmo não constituindo, sozinha, um fator suficiente para diferenciar situações.

Um problema que nos colocamos é como definir os conceitos de contexto público e contexto particular. As decisões tomadas no C-ORAL-ROM não parecem completamente coerentes entre os vários grupos. Naturalmente mais de uma decisão podia, legitimamente, ser tomada. O que nos interessa aqui é explicitar os critérios usados no nosso *corpus*. A definição de contexto público versus contexto particular/familiar deve ser entendida com base no papel exercido pelo falante no momento da interação. Um determinado falante pode interagir na sua própria condição de indivíduo, como acontece normalmente quando se interage com amigos ou com familiares; ou pode agir com base em um papel social, seja devido ao trabalho ou seja devido à predominância, naquela interação, de uma condição específica determinada pelas condições de natureza social ou pelo assunto e não pelo relacionamento entre dois ou mais indivíduos específicos.

Portanto, não é a presença ou a ausência de estranhos no contexto de interação que determina necessariamente, na nossa escolha, o caráter público ou particular da situação. Uma

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 20-35 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

conversa em um restaurante entre amigos íntimos não pode ser automaticamente considerada pública, assim como não pode ser considerada particular uma interação profissional entre um aluno e seu orientador, mesmo se os dois estão sozinhos no escritório do orientador. É evidente que alguns fatores favorecem a criação de uma situação que consideramos pública ou particular, mas nenhum fator é decisivo em si mesmo. O que conta é o comportamento do falante, independentemente do motivo que o leve a esse comportamento: se ele se comporta com base no papel social (cliente, profissional, funcionário, cidadão) consideramos a interação pública; se o comportamento é com base na própria identidade individual consideramos a interação particular. Alguns exemplos podem ilustrar melhor os critérios decisórios: em um caso gravamos uma entrevista com a proprietária de um restaurante, conduzida pela irmã da mesma. A entrevista aconteceu no restaurante. As perguntas eram relativas à condução do estabelecimento. O que nós esperávamos era um comportamento de natureza familiar, dada a relação entre as duas interlocutoras. Escutando depois a entrevista, notamos, sem sombra de dúvida, que a entrevistada manteve por toda a entrevista um comportamento absolutamente diferente daquele que normalmente mantém com a irmã, e se comportou com base no papel de proprietária de restaurante que ilustra, mesmo se informalmente, as características da própria profissão. O lugar da entrevista e o assunto foram nesse caso mais importantes que a relação com o interlocutor para determinar o comportamento do falante. Ao contrário, temos uma gravação em que um grupo de alunos universitários conversa ao lado do elevador da faculdade, expostos à passagem de outros alunos, professores, funcionários e estranhos em geral. O conteúdo e o tom da conversa foram claramente de tipo particular; o comportamento dos alunos não foi diferente daquele que teriam tido se a reunião tivesse sido na casa de um deles.

4. O cabeçalho

Os metadados presentes no cabeçalho seguem exatamente o mesmo critério do C-ORAL-ROM, com a única diferença, já vista, quanto à definição das faixas de escolaridade. Portanto, um exemplo de cabeçalho é o seguinte:

@Title: Daughter
@File: ifammn06
@Participants: CAR, Carmosina (woman, C, 1, housekeeper, narrator, Alpercata (MG))
MAR, Maryualê (woman, B, 3, professor, intervenient, Florianópolis)
@Date: 12/04/2008
@Place: Belo Horizonte
@Situation: narration about how CAR adopted her youngest daughter, CAR's kitchen, CAR makes lunch, not hidden, researcher participant (CAR works as housekeeper at the researcher's home)
@Topic: daughter's adoption
@Source: C-ORAL-BRASIL
@Class: informal, familiar/ private, monologue
@Length: 9’51”
@Words: 1508
@Acoustic quality: A
@Transcriber: Maryualê M. Mittmann
@Revisor: Heloisa P. Vale

@Comments: text collected and recorded by Maryualê M. Mittmann. CAR pronounces "dócia" and "vivendos" when it should be "dócil" and "vivendo". Sometimes CAR calls the researcher Mara and not Mary.

A partir dos dados armazenados no cabeçalho será possível fornecer os dados estatísticos sobre as categorias de sexo, idade, nível escolar, ocupação e papel comunicativo exercido pelos falantes, assim como fornecer informações sobre as atividades exercidas durante a gravação, os locais de gravação e os assuntos. Os comentários, que nesse exemplo são pequenos, hospedam normalmente observações sobre a natureza situacional, julgadas importantes para a compreensão do texto como um todo ou de suas partes, e observações lingüísticas não acolhidas nos critérios de transcrição, mas julgadas relevantes, no caso específico. A intenção é usar o espaço para comentários para reduzir ao mínimo o recurso às linhas dependentes previstas pelo CHILDES-CLAN. A razão disto é que as linhas dependentes dificultam a leitura do texto e geralmente são ignoradas, em corpora dessa natureza. Sua utilidade é estritamente ligada à função original do sistema, que era a de dar conta da linguagem infantil, que necessita de um acompanhamento comentado constante para sua correta interpretação.

5. As modalidades de gravação

Com poucas exceções, as gravações são realizadas em formato .wav com o seguinte equipamento:

- gravadores digitais Marantz PDD660, com cartão de memória Compact Flash de 2 gigabytes;
- kits wireless Sennheiser Evolution EW100 G2 (receiver, transmitter, microfone de lapela), com dois kits bateria/carregador adaptados para o receiver, ou solução nativa com bateria própria e seis microfones completos;
- microfones omnidirecionais Sennheiser MD 421, com pedestais Hunter PMP103, cabos RCL303569 de 6 metros, ou sistema wireless;
- mixer Xenyx 1222, com cabos para seis entradas de microfones de lapela.

Os microfones de lapela são utilizados tanto para situações dialógicas quanto monológicas, enquanto o microfone omnidirecional é utilizado para conversações. Em um segundo momento, por causa da dificuldade de se obter boas gravações com o microfone omnidirecional, resolvemos gravar as conversações com microfones de lapela e mixer, de maneira que até seis falantes podiam ser gravados com microfones monodirecionais sem fio. Contudo, em alguns casos obtivemos boas gravações de conversações a três (ou no máximo quatro) participantes com dois microfones de lapela, quando os falantes sem microfones estavam posicionados perto de um falante com microfone, e não faltam também boas gravações obtidas com o microfone omnidirecional.

Esse equipamento permite gravações em ambiente natural com uma qualidade de som que, no mínimo, permite recuperar a curva de F0, fundamental para as análises prosódicas a serem executadas com o software WinPitch; porém, em geral, a qualidade obtida é apropriada também para análises segmentais.

Em geral, quando realizadas com os microfones de lapela, as gravações são mantidas em estéreo. Isso oferece maiores recursos para a análise da fala: por exemplo, permite identificar melhor a fala de cada falante em caso de sobreposição, além de gerar uma qualidade melhor do som. As poucas exceções ao uso desse equipamento são devidas a um número muito restrito de

gravações efetuadas em cabine acústica ou em formato .mp3. As gravações estão sendo coletadas com informantes que se disponibilizem a assinar o termo de consentimento, autorizando a sua gravação, aprovado pelo Comitê de Ética em Pesquisa da UFMG[11].

A qualidade acústica das gravações é indicada no cabeçalho de cada transcrição. O critério seguido para indicar a qualidade acústica difere daquele do corpus C-ORAL-ROM por causa das inovações tecnológicas acontecidas desde que as gravações do mesmo começaram (algumas datam até dos anos 1980). As gravações do C-ORAL-BRASIL foram iniciadas somente em 2007. O C-ORAL-ROM usa as letras A, B e C para indicar, respectivamente, qualidade excelente, boa e aceitável. Na coleta dos dados a letra D indica que a qualidade acústica do arquivo é ruim demais para que faça parte do corpus. No C-ORAL-BRASIL essas letras foram mantidas, porém com duas diferenças:

1. em princípio, a mesma letra indica uma qualidade melhor no C-ORAL-BRASIL do que aquela indicada pela mesma letra no C-ORAL-ROM, devido às melhoras tecnológicas;
2. no C-ORAL-BRASIL adotamos uma codificação da qualidade acústica mais complexa, usando, além das três possibilidades A, B e C, também os pares AB, BC, CD para qualidades intermediárias.

A indicação da qualidade acústica depende dos seguintes parâmetros: a possibilidade de escutar claramente a fala dos informantes; qualidade do espectrograma; presença ou falta de ruídos; presença ou falta de retorno; ganho; sobreposições e a situação de gravação. A qualidade A indica gravações que, em princípio, permitiriam muitos tipos de estudos fonéticos; ao contrário, D indica gravações que podem ser usadas somente para estudos de natureza morfossintática ou lexical, porque na maioria dos casos a F0 é ilegível ou inconfiável. Os códigos intermediários garantem no mínimo a confiabilidade e legibilidade de pelo menos 60% da curva. A codificação CD indica que as gravações assim codificadas podem ser utilizadas no corpus, porém somente nos casos em que não devem ser substituídas, por razões situacionais. Naturalmente, o rigor da classificação deve considerar também o tipo de interação e a situação na qual ela acontece. Os critérios são mais exigentes para os monólogos, um pouco menos para os diálogos e ainda menos para as conversações. O simples fato de que as conversações possuem um número mais alto de falantes condiciona negativamente, por si só, a qualidade acústica. Analogamente, não se pode esperar que um diálogo gravado em um supermercado tenha a mesma qualidade de uma interação gravada em um lugar silencioso. Portanto a conclusão é que a codificação indica a qualidade por si mesma, mas é possível que uma conversação A, dados os parâmetros expostos, corresponda a um monólogo AB.

6. As modalidades de transcrição

As transcrições são realizadas segundo o sistema CHILDES-CLAN (MACWHINNEY 2000), implementado para a notação prosódica com as indicações de Moneglia & Cresti (1997). A transcrição é de base ortográfica, mas várias adaptações foram introduzidas para capturar alguns fenômenos da fala do PB considerados especialmente importantes. Os critérios de transcrição serão discutidos sistematicamente em outra ocasião. Aqui damos as diretrizes fundamentais e alguns exemplos indicativos. A transcrição tenciona capturar fenômenos que podem refletir

[11] COEP

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 20-35 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

fenômenos de gramaticalização e lexicalização em curso. Ao mesmo tempo, os critérios devem considerar duas outras circunstâncias: em primeiro lugar, o texto transcrito não pode gerar problemas para a imediata compreensão do leitor; em segundo lugar, é necessário que as formas não ortográficas sejam ligadas a critérios que garantam a homogeneidade dos transcritores, e não gerem mais confusão que vantagens. Advém daí a dificuldade de encontrar um equilíbrio entre a necessidade de capturar alguns fenômenos, a legibilidade do texto e a factibilidade da transcrição executada por muitos transcritores (o grupo base é composto por sete pessoas). Alem de critérios mais gerais, como a transrcição de números, siglas, letras do alfabeto, palavras estrangeiras, etc., alguns dos fenômenos que a transcrição pretende capturar são:

- formas aferéticas: as formas dos verbos *estar* (*tô, tá, tando, tar*, etc.) e de outros verbos (*agüentar > güentar*; *espera > pêra*; etc.) ou formas não verbais (*obrigado > brigado*);
- formas apocopadas: *pode fazer > po' fazer*; *olha > o'*; etc;
- ausência de marca de plural nos elementos não iniciais do sintagma nominal. A marca de plural no PB é freqüentemente omitida quando ela já está presente no determinante e às vezes até em um elemento invariável: *os meninos bonito > os menino bonito*; *ques menino bonito!*, etc;
- os fenômenos de cliticização dos pronomes sujeitos: *você(s) > cê(s)*; *ele > e'*; *ela > ea*; *eles*; *> es*; *elas > eas*. Devido à maior variedade de realização do pronome de primeira pessoa, sua percepção oferece maior discordância entre os transcritores o que torna a sua representação menos confiável;
- as formas do tipo *de´/des/dea/deas* para *dele/deles/dela/delas* e as equivalentes com outras preposições, como em *pr´ es, c´ es, d´ es*, ou dos demonstrativos *aques* para *aqueles* e suas respectivas formas preposicionadas;
- os fenômenos de cliticização da negação preverbal *nũ*;
- redução das preposições: *para/pra/pa/p´*; *com/c´*; *de/d´*; *em/ni/n´*;
- contração das preposições articuladas: *para* + art. *> pro, pra, pros, pras, prum*; de + art. indefinidos: *dum, duns*; com + art.: *co, ca, cos, cas*, etc;
- alomorfia dos diminutivos masculinos: *sozinho > sozim*;
- algumas exclamações: *Nossa/No'/Nu´*; *Vixe'/Vix´* (*Virgem Maria*), etc;
- redução da morfologia verbal: os tipos *eles faz*; *eles foi*, etc;
- as formas de primeira pessoa plural verbal não padrão: os tipos *empurramo* / *empurremo*;
- as formas de subjuntivo não padrão: o tipo *seje* (por *seja*);
- a forma *tem* e *tem que* tanto nas perífrases *eu tenho que* + infinitivo quanto no uso de *tenho* + objeto na primeira pessoa: *eu tem que sair agora* e *eu não tem vontade de sair agora*;
- a distinção entre as construções aspectuais do tipo *ele pegou e falou*; *ele foi e fez* e, por outro lado, as construções com verbos seriais, do tipo *ele pegou falou*; *ele foi fez*, sem o uso da conjunção;
- as formas com a perda do verbo *ser* em estruturas interrogativas focalizadas, do tipo: *que que cê faz*; *quando que cê vem*; *porque que cê diz isso*, etc;
- a queda do relativo em várias construções, como em *isso cê tá fazendo é bom* para *isso que cê ta fazendo é bom*; - etc.

Os exemplos ilustram somente uma parte dos fenômenos que os critérios pretendem capturar. Mas, mesmo assim, é evidente que critérios dessa natureza apresentam dificuldades e ao mesmo tempo abrem novas perspectivas de estudo para fenômenos conhecidos, porém nunca estudados, com métodos quantitativos e com a possibilidade de verificar estatisticamente sua co-ocorrência. O principal problema acarretado pelos critérios de transcrição é relativo à

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 20-35 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

necessidade de programar um etiquetador léxico-morfossintático capaz de identificar formas, com um baixo grau de erro. Esse problema se torna maior se consideramos que a unidade de base do corpus não é a sentença mas o enunciado, e que as regras da escrita não ajudam a prever com base em distribuição as funções morfossintáticas das várias formas. Para superar esse problema contaremos com a parceria do autor do analisador sintático PALAVRAS (BICK 2000-2002), o software para o português que deu os melhores resultados em testes com nossas transcrições.

Contudo, consideramos que o esforço vale à pena. De fato uma transcrição exclusivamente ortográfica (como aquela do corpus de português do C-ORAL-ROM) esconde definitivamente os fenômenos mais interessantes da fala a nível léxico-morfossintático. Para o corpus italiano a escolha feita foi diferente, e se procurou mostrar, através de uma atenta escolha dos critérios de transcrição, os principais fenômenos de lexicalização e gramaticalização em curso. Todavia, se essa operação é em si difícil, é mais árdua ainda se pensamos nas características do PB e, em particular, naquelas da variedade de Minas Gerais. Uma coisa é transcrever os fenômenos de alomorfia de uma língua de base silábica como o italiano, com poucos e limitados fenômenos de coarticulação e com uma tradição ortográfica antiga para fenômenos locais, quase todos relativamente estabilizados; outra, completamente diferente, é perseguir o mesmo objetivo com uma língua em rápida transformação e cuja alomorfia se manifesta com um nível de coarticulação altíssimo devido ao forte componente acentual.

Mas, se o esforço é significativo e requer uma formação atenta dos transcritores, além de um processo longo de revisão, dispor de transcrições desse tipo nos permitirá desenvolver estudos interessantíssimos, com base estatística, para observar a co-ocorrência de vários fenômenos, tanto dentro de cada ramificação do corpus quando a nível diastrático e individual.

7. As modalidades de segmentação

As transcrições são segmentadas em enunciados e unidades tonais, segundo os mesmos critérios adotados para o C-ORAL-ROM, com poucas e pequenas mudanças. Os símbolos usados são:

- a barra dupla (//) indica quebra entonacional peceptível de valor terminal, e, portanto, fronteira de enunciado; o enunciado é definido, com base na Teoria da Língua em Ato, como a menor unidade autônoma pragmaticamente, dotada de perfil entonacional perceptível como terminal (t´HART-COHEN-COLLIER 1990), ou seja, a contraparte lingüística do ato de fala;
- a barra simples (/) indica quebra entonacional perceptível de valor não terminal, e, portanto, fronteira de unidade tonal (em princípio com valor informacional) dentro de enunciado;
- o signo (+) indica enunciado interrompido. Portanto possui valor terminal, mas sinaliza que o enunciado não foi completado, qualquer que seja a razão disso;
- o signo ([/n]) indica retracting. O número ao lado da barra indica o número de palavras envolvidas no retracting e cancelada pelo falante.

Damos, a seguir, exemplos de quebras em fragmentos de fala extraídos do *corpus*.

(1) Seguem alguns casos de mais enunciados simples no mesmo turno, cada enunciado apresentando somente quebra terminal:

*PAU: bom // Rogério //
*FLA: é // me falaram // que ele é muito <bom> //
*FLA: hhh o nosso tá longe // tá em outra cidade //
*REN: ham ham // <é> // tá // tá certo // obrigada //

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 20-35 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

(2) Seguem alguns casos de enunciados complexos, com quebras não terminais em seu interior. Em cada exemplo identificamos a unidade de comentário, ou seja, a única que possui autonomia pragmática:
*FBA: Tudo que é de bom / pra gente / que a gente tá se sentindo que realmente tá fazendo / né / e [/] e que tá dando retorno / a gente continua //=COM=
(3) Um exemplo de enunciado interrompido, seguido pela interrupção do outro falante:
*PAU: aí / por exemplo +
*ROG: aqui já tá dando [/4] aqui já tá dando a altura //
(4) Um exemplo de retracting:
*ROG: aqui já tá dando [/4] aqui já tá dando a altura //
(5) Um trecho de fala transcrito e segmentado:
*PAU: bom // Rogério //
*ROG: hum //
*PAU: cê sabe aqui como é que [/3] como é que tem que fazer esse muro aqui / né // por que que cê não tá trabalhando com linha aqui / o' //

Segue, na figura 1, um detalhe do espectrograma (com o uso do software WinPitch de Philippe Martin[12]) relativo à gravação exemplificada. Trata-se de uma gravação ao ar livre entre um engenheiro e um pedreiro durante a construção de um muro. A qualidade acústica é AB.

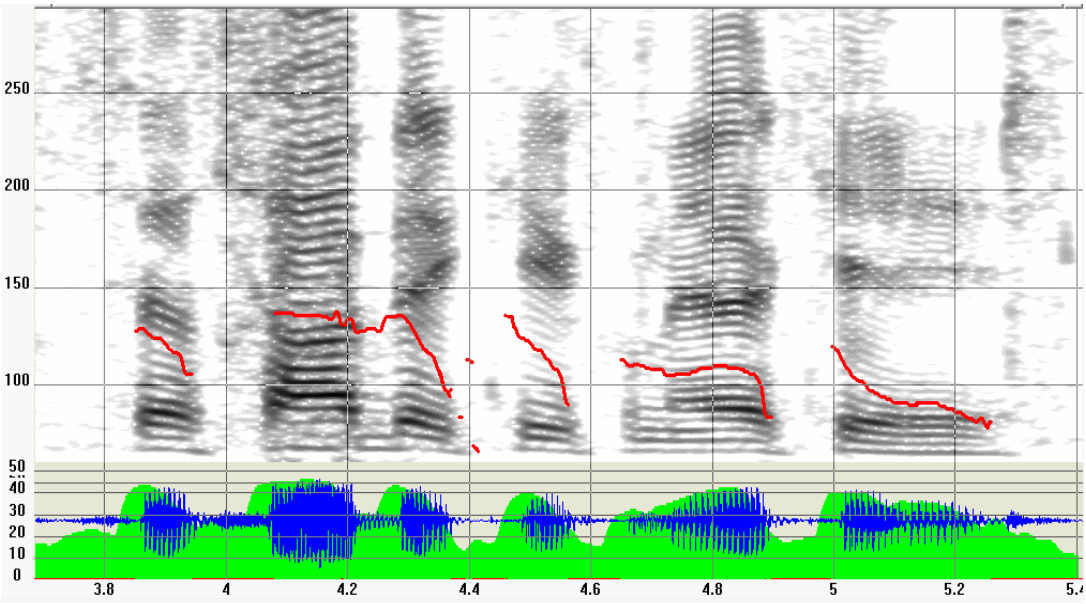

Figura 1

8. O processo de transcrição e segmentação: a validação

O processo de transcrição e segmentação é bastante complexo[13]. Um longo treinamento foi feito durante vários estágios, que incluem três workshops e uma disciplina de pós-graduação. Todos

[12] www.winpitch.com
[13] Para os protocolos seguidos no C-ORAL-ROM, vejam-se Danieli *et alii* (2004); Moneglia *et alii* (s/d).

os participantes executaram transcrições e segmentações. É importante marcar que a segmentação é contemporânea à transcrição, em se tratando ambas de atividades fruto da percepção acústica. Depois desse treinamento, um grupo de dez potenciais transcritores foi selecionado para alguns testes. Os resultados detalhados dos testes e do processo de formação serão expostos em outra ocasião. Aqui oferecemos apenas as indicações principais.

Depois do treinamento, os dez transcritores potenciais foram divididos em três grupos, dois de três e um de quatro transcritores. A divisão em grupos foi feita segundo os critérios seguintes: o grupo 1 foi formado com os três alunos que tinham mostrado mais aptidão no processo de segmentação durante o treinamento e ao mesmo tempo maior comprometimento com projeto. Tratava-se de uma aluna de doutorado, uma de mestrado e uma bolsista de iniciação científica. O grupo 2 foi formado com as três alunas, dentre os restastes, que tinham em princípio um nível de preparo e comprometimento maior: uma aluna de doutorado, uma mestre que acabou entrando no programa de doutorado e uma bolsista de graduação. O grupo três foi formado com pessoas que tinham entrado mais recentemente no projeto ou tinham um grau de comprometimento menor: 3 alunos de graduação (dois bolsistas) e uma mestranda. Essa última não está integrada ao projeto, mas havia acompanhado o treinamento e demonstrado interesse em participar do processo. Os três alunos de graduação tinham menos experiência de segmentação, mas por isso mesmo tinham, em princípio, maiores chances de melhorar o seu desempenho.

Cada grupo foi submetido a uma série de testes, depois dos quais, em reuniões coletivas, se discutiam os casos de desacordo na marcação das quebras. Os grupos dois e três foram mudados depois dos primeiros testes. A razão é a seguinte. Nós não precisávamos de três grupos de transcritores, mas somente de dois. Estava, portanto, nos planos o fato de que, se o grupo 1 tinha muitas chances de se revelar o mais confiável, os componentes dos outros dois grupos podiam revelar, ao longo dos testes e das discussões, um desempenho diferente daquele inicial. Na verdade, o grupo 2 mostrou resultados sempre um pouco melhores do que o grupo três, mas dois dos componentes do grupo 3 mostraram grande comprometimento e melhoras rápidas. Por isso foi decidido que o segundo grupo fosse formado novamente, mudando dois de seus componentes e adicionando um terceiro. O grupo dois passou, portanto, a ser de quatro componentes. Atualmente o grupo alcançou um kappa a quatro de 0,82 (0,85 se consideramos somente os três melhores), mas para isso levamos de cerca de dois meses de treinamento a mais do que para o grupo 1. Uma outra informação importante é relativa ao fato de que as transcrições usadas para os testes eram transcrições provisórias, ainda não aperfeiçoadas quanto a todos os critérios. Isso gerou desacordos não devidos a problemas de segmentação, mas à falta de um critério completamente homogêneo para decidir vários casos que em princípio podiam ter soluções diferentes. Ou seja, o grau de acordo teve uma pequena diminuição por motivos não ligados à competência na segmentação.

Damos aqui os resultados dos testes somente para o grupo 1:

1. Segmentação de cerca de 800 palavras de um texto dialógico e 800 palavras de um texto monológico. Teste kappa: 0,820 para o dialógico; 0,750 para o monológico.
2. Segmentação de cerca de 1500 palavras de um texto dialógico. Teste kappa: 0,820.
3. Segmentação de cerca de 1500 palavras de um texto monológico. Teste kappa: 0,839.

Esses resultados precisam de alguns comentários mais qualitativos. Em primeiro lugar, era um objetivo importante não ter nenhum desacordo quanto às quebras terminais. Em todos os testes alguns poucos desacordos (entre sete, no pior caso, e dois no melhor caso) aconteceram. Nós conferimos as instâncias de desacordo e sempre verificamos ter-se tratado ou de simples

distração, imediatamente corrigida pelo segmentador na hora da averiguação, ou, mais freqüentemente, de casos motivados por questões ainda não resolvidas em fase de transcrição. Um exemplo típico dessa modalidade de problema é representado pelos turnos ocupados somente por riso. Em alguns casos o segmentador achou que o turno ocupado somente por riso não devesse ser segmentado; em outro achou que devesse ser segmentado com quebra terminal. Isso gerou a maior parte dos poucos desacordos extremos entre quebra terminal e falta de quebra. Somando as distrações e os casos de diferentes comportamentos em casos ambíguos, verificamos que não existia, de fato, nenhum desacordo que pudesse gerar uma contraposição entre quebras terminais e nenhuma quebra.

Em segundo lugar, a averiguação em todos os testes levou a uma imediata redução dos desacordos em fase de discussão. Na maioria dos casos, o que acontecia era que o segmentador que tinha escolhido a posição minoritária, mesmo sem saber que a posição dele era a minoritária, ou não reconhecia que tinha escolhido aquela posição ou mudava de posição imediatamente. Esses dois fatores geram uma melhora clara nos resultados que já seriam excelentes do ponto de vista estatístico, como já documentado.

Depois dessa fase, nós consideramos que a consistência dos resultados garantia uma base satisfatória, mas, para chegar a resultados de excelência, resolvemos continuar os testes diferenciando as quebras terminais das não terminais. Fizemos, portanto, primeiro um teste somente relativo às quebras terminais. O resultado foi de 0.901, que seria ainda superior se tirássemos os poucos casos de desacordo devido às ambigüidades normais. Depois fizemos um teste relativo às quebras não terminais. Nesse caso o teste deu um resultado de 0,660. Esse resultado, somado ao teste para as quebras terminais ainda dava um êxito claramente superior ao 0,8, considerado excelente, mas nos induziu a buscar com maior profundidade as diferenças. Resultou claro que, mesmo desconsiderando os erros de distração, imediatamente desmentidos em fase de averiguação, uma das três segmentadoras tinha uma tendência evidente em perceber as quebras de maneira mais fraca. Isso tem pouco impacto quanto às quebras terminais, mas possui um impacto mais evidente nas quebras não terminais. Decidimos, portanto, que a melhor maneira para começar as transcrições seria deixar a fase de revisão somente para as duas transcritoras que apresentam o maior grau de acordo. Mesmo assim, resolvemos continuar o processo de discussão e de testes por mais tempo.

Enfim, mesmo antes da revisão e com transcrições ainda aproximadas e que, portanto, geram ambigüidade, podemos contar com um kappa de acordo superior a 0,8, ou seja já excelente. Esse kappa será automaticamente melhorado na hora em que as transcrições forem definitivas e será muito melhorado com as revisões. No momento, o grupo 1 acabou as segmentação dos textos que deverão compor o minicorpus com os melhores textos (cerca de 20% do corpus inteiro), enquanto os transcritores do grupo 2, chegaram a transcrever cerca de 50% do restante do corpus. Em novas reuniões estamos verificando a segmentação do grupo 1, antes que comece o processo de revisão. Mesmo sem testes é raríssimo encontrar desacordo quanto às quebras terminais e poucos desacordos quanto às quebras não terminais. Não registramos nenhum caso de desacordo extremo. Depois das revisões novos testes deverão confirmar um acordo ainda maior.

Vale à pena notar que a menor sensibilidade quanto às quebras em uma das transcritoras (o que não permitiu alcançar um kappa ainda mais alto) pode ser devida a uma causa interessante: essa transcritora é originária de uma área rural, enquanto as duas transcritoras são originárias de áreas urbanas. A fala de área rural, principalmente em Minas Gerais, parece ser caracterizada por um andamento mais acentual do que a fala urbana. A fala acentual apresenta

características rítmicas que reduzem a percepção das quebras mais fracas, as quais geralmente são devidas a fenômenos de escansionamento e que não refletem padrões informacionais. Esse tipo de sensibilidade diferente, ligada a uma fala mais acentual merece ser estudada no futu

9. Realização de um minicorpus para estudos

Para garantir a disponibilização de material necessário para estudos com agilidade, resolveu-se segmentar o trabalho da equipe em duas frentes paralelas. Por um lado transcrição, segmentação, alinhamento e etiquetagem informacional de um minicorpus balanceado. Este consiste de 30.000 palavras e cerca de 5000 enunciados para assegurar a presença de pelo menos 2000 enunciados complexos (cerca de 30 textos, considerado que alguns monólogos são menores de 1500 palavras). Por outro lado transcrição, segmentação e alinhamento dos outros textos restantes. O minicorpus será completamente etiquetado informacionalmente. Atualmente, com a transcrição completada, paralelamente à revisão e depois ao alinhamento, o grupo 1 está passando por um processo de formação para a etiquetagem informacional.

A realização do minicorpus obedece a critérios de máxima qualidade. Participam do minicorpus, seguindo as várias ramificações, somente os textos que oferecem a melhor combinação possível com base nos seguintes parâmetros:

- representatividade da ramificação. O texto deve ser um bom protótipo do tipo de variação (público x familiar/particular; monológico x dialógico x conversacional);
- maior variação possível de atividade. Nunca dois textos com a mesma situação comunicativa;
- alta qualidade acústica. A qualidade acústica é dada pelos seguintes fatores: qualidade do espectrograma; pouco ou nenhum ruído de fundo; pouco ou nenhum retorno do sinal; claridade da voz; ganho bom; cálculo confiável da F0; porcentagem de sobreposição baixa;
- diversidade dos locutores. Nunca um locutor pode aparecer duas vezes, a não ser no caso em queele apareça como interlocutor de um monólogo. Busca-se também uma representatividade tendencialmente paritária de vozes masculinas e femininas, e de falantes de idades variadas;
- não marcação diastrática. Busca-se de preferência textos que não sejam de diastratia muito baixa nem muito alta.
- interesse do conteúdo. O interesse do conteúdo é um valor em si por duas razões: aumenta a atenção de transcritores e segmentadores; aumenta o nível de informatividade, pois garante uma fala mais espontânea.

Sempre para que o minicorpus alcance o nível máximo de qualidade, os textos escolhidos para integrá-lo são de exclusiva responsabilidade do grupo 1, com a revisão somente dos dois transcritores do grupo que alcançaram o acordo máximo.

10. O estágio atual e os próximos passos

Atualmente temos mais de 180 gravações, que já permitiriam cobrir todas as ramificações do *corpus*. Como muitas gravações são longas, em alguns casos com duração de até várias horas, é certo que em várias delas é possível conseguir mais de um texto. O número de textos disponíveis é sem dúvida suficiente para o *corpus*. De toda maneira, o processo de gravação não será nunca considerado concluído, procurando-se sempre o acréscimo de mais textos a fim de permitir-se uma escolha maior e de melhor qualidade para o *corpus*. A ampliação do corpus para além dos objetivos do projeto C-ORAL-BRASIL é útil para expandir a gama de fenômenos

passíveis de serem estudados de maneira confiável (seguindo o exemplo do corpus italiano LABLITA[14] que em muito supera as dimensões do corpus italiano do C-ORAL-ROM).

O programa prevê que, até junho de 2009, tanto o minicorpus (transcrição, revisão e alinhamento) estará completo, como também a transcrição do resto do *corpus*. Até junho de 2009 também deverá ser completada a formação para o alinhamento do minicorpus. Até dezembro 2009 o *corpus*, em sua integridade, deverá estar transcrito, revisto e alinhado.

Com a revisão das transcrições do minicorpus será possível começar o treinamento e a programação do etiquetador PALAVRAS a fim de que se obtenha uma versão do mesmo apropriada para *PoS tagging* do *corpus* inteiro.

A mesma macro e a mesma folha Excel usada para o C-ORAL-ROM fornecerão as principais medidas da fala, junto ao *software* usado para o *PoS tagging*: para cada tipologia interacional (monólogo, diálogo e conversação) serão calculados o número de turnos, enunciados e unidade tonais em relação ao tempo e ao número de palavras; o número de enunciados interrompidos e de fenômenos de *retracting*; o número de enunciados com e sem verbo, cruzando esse dado com a constituição simples ou complexa do enunciado; a freqüência de ocorrência da negação e das conjunções *E*, *MAS*, *PORQUE* e *QUE*, e sua posição (começo de turno, começo de enunciado, começo de unidade tonal, dentro de unidade tonal, unidade dedicada, ou seja, quando a unidade é ocupada somente pelo item investigado)[15].

Essas mesmas medidas já estão sendo investigadas de forma qualitativa em um grupo de nove textos (três diálogos, três monólogos e três conversações) que participarão do minicorpus. O estudo qualitativo, de fato, permite observar fenômenos que fogem a um inquérito quantitavo com instrumentos computacionais. Por exemplo, o etiquetador informático nos diz se um enunciado possui um elemento verbal, mas não nos diz se esse elemento é verbal somente do ponto de vista morfológico ou também funcionalmente; assim, itens como *tá* ou *sei* são etiquetados da mesma forma tanto se aparecem em enunciados como *o meu amigo tá bem* ou *eu sei o que estou dizendo* quanto se aparecem em enunciados em que *tá* e *sei* valem como *sim* ou *ok*. Analogamente o etiquetador não captura o valor funcionalmente não verbal da pergunta eco do verbo para afirmar ou concordar, como no uso freqüente no PB[16]. Ademais, uma análise qualitativa permite distinguir os enunciados verbais em que o verbo constitui realmente o núcleo do enunciado daquele em que o verbo aparece em posição não nuclear. Um outro exemplo do quanto uma análise qualitativa pode ser importante é relativo à freqüência de ocorrência de enunciados não verbais caracterizados unicamente por expressões como *hum hum* ou *ahn ahn*, lingüisticamente vazias mesmo se convencionalizadas como afirmações ou negações.

Concluímos esse trabalho expondo um dos desdobramentos do projeto que nos parece mais interessante, pelas grandes potencialidades descritivas, mas principalmente pela possibilidade que oferece de responder a uma pergunta base do projeto C-ORAL-ROM: o que é próprio da fala e o que é específico de uma dada língua/cultura. Estamos nos referindo à possibilidade, através do *corpus* C-ORAL-BRASIL, de comparar a estrutura da fala espontânea entre o PB e o português europeu (PE) com base em parâmetros não somente segmentais,

[14] http://lablita.dit.unifi.it/corpora/

[15] Para alguns estudos sobre as línguas incluídas no C-ORAL-ROM, veja-se Moneglia (2004, 2006); Cresti-Moneglia (2007).

[16] Por exemplo no fragment dialógico seguinte: *XYZ: o João viajou pra Itália // *XYZ: viajou //. O segundo enunciado nesse caso vale como um *sim*.

morfossintáticos e lexicais, mas também prosódicos, informacionais e ilocucionários. Isso nos permitirá estudos para tentar responder a duas grandes questões:
1. Até que ponto o PE e PB, duas variantes da mesma língua, se estruturam da mesma maneira e até que ponto ele se estruturam com base em parâmetros diferentes, talvez ligados às suas diferentes matrizes culturais. Os estudos do grupo LABLITA sobre o italiano constituirão automaticamente um metro de comparação. Já os primeiros estudos sobre o PB, com base no projeto piloto, notaram diferenças significativas na estruturação informacional entre italiano e PB. Saber se o PE segue mais a estruturação de uma língua diferente, mas da mesma cultura (italiano) ou de uma variante culturalmente distante da mesma língua (PB) poderá nos ajudar a entender muito sobre a estruturação da fala. Por isso um dos próximos passos do projeto será etiquetar um minicorpus do *corpus* de PE presente no C-ORAL-ROM para começarmos estudos comparativos entre três línguas: italiano, PE e PB.
2. A comparação entre três línguas (italiano, PE e PB), considerado que o PE compartilha o código de base com uma das duas e a matriz cultural com a outra, permitiria isolar os fenômenos do PB que se apresentam como candidatos naturais a uma explicação em termos de contato lingüístico. Se em alguns aspectos da fala, principalmente nos aspectos entoacionais e informacionais, o PE revelar maior semelhança com o italiano do que com o PB, poderíamos com boas bases estudar no PB os mesmos aspectos buscando uma explicação na história de contato com línguas de origens muito diferentes.

ABSTRACT: This paper reports the main aspects related to the compilation of a spontaneous speech corpus of Brazilian Portuguese, the C-ORAL-BRASIL, focusing on its textual typology, transcription and segmentation criteria, the methodological procedures, the transcribers'training as well as the very first descriptive results emerging from a pilot study.

Keywords: Spontaneous speech corpus, Brazilian Portuguese; Utterances; Information units.

Bibliografia

ALVES DE DEUS, L. *O Tópico no português do Brasil*. Dissertação de Mestrado, UFMG, 2008 (Orientador: Tommaso Raso).
BICK, E. *The Parsing System Palavras – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press, (2000-2)
CRESTI, E. *Corpus di italiano parlato*. Firenze: Accademia della Crusca, 2000, 2 vol. CRESTI E. Per una nuova classificazione dell'illocuzione. In: E. Burr (ed.), *Tradizione e innovazione* - Atti del VI convergno SILFI (Duisburg 28.06/02.07 2000). Firenze: Cesati, 2005, pp. 233-246.
CRESTI E. Illocuzione e modalità. In: BECCARIA, P. , MARELLO, C. (eds.). *La parola al testo. Scritti per Bice Mortara-Garavelli*. Torino: Ed. dell'Orso, 2002, pp. 133-145.
CRESTI E. Illocution et modalité dans le comment et le topic. In A. SCARANO (ed.) *Macro-syntaxe et pragmatique. L'analyse linguistique de l'oral*. Roma: Bulzoni, 2003.
CRESTI, E., BACELAR do Nascimento, F., MORENO Sandoval, A., VERONIS, J. , MARTIN, Ph., CHOUKRI, K. The C-ORAL-ROM CORPUS. A multilingual resource of spontaneous speech for romance languages. In: LINO, M.T., XAVIER, M.F., FERREIRA, F., COSTA, R., SILVA, R. (eds.) *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Paris: ELRA, 2004, vol. 2, pp. 575-79.

CRESTI, E. – MONEGLIA, M. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins, 2005.
CRESTI, E. - MONEGLIA, M. C-ORAL-ROM. Comparing Romance Languages in Spontaneous Speech Corpora. In: SILVA, T. C. e MELLO, H. (eds.). *Conferências do V Congresso Internacional da Associação Brasileira de Lingüística*. Belo Horizonte: UFMG, 2007.
t´HART, J. – COHEN, A. – COLLIER, R. *A perceptual study on intonation*: *an experimental approach to speech melody*. Cambridge: Cambridge University Press, 1990.
MACWHINNEY, B. J. *The CHILDES Project. Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum, 2 vol., 2000.
MAIA ROCHA, B., RASO, T., ANDRADE, M. I.. Alguns auxílios dialógicos em italiano, português do Brasil e em italianos cultos em contato prolongado com o português do Brasil. In: *Fragmentos*, in press.
MARTIN, Ph., WinPitch (www.winpitch.com).
MONEGLIA, M. Specifications on the C-ORAL-ROM Corpus. http://lablita.dit.unifi.it/coralrom/papers/Specifications-CORALROM.pdf, 2000.
MONEGLIA, M. Measurements of Spoken Language Variability in a Multilingual Corpus. Predictable Aspects. In: : LINO, M.T., XAVIER, M.F., FERREIRA, F., COSTA, R., SILVA, R.(eds.). *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Paris: ELRA, 2004, pp. 1419-22.
MONEGLIA, M. C-ORAL-ROM. Un corpus di riferimento del parlato spontaneo per l'italiano e le lingue romanze. In: KORZEN, J. (ed.). *Lingua, cultura e intercultura. L'italiano e le altre lingue*. Atti del VIII convegno SILFI (Copenhagen 22-26 July 2004). Copenhagen: Samfunzliteratur Press, 2005, pp. 229-42.
MONEGLIA, M. Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective. In: KAWAGUCHI, Y., ZAIMA, S., TAKAGAKI, T. (eds.). *Spoken Language Corpus and Linguistics Informatics*. Amsterdam-Philadelphia: John Benjamins, 2006, pp. 153-79.
MONEGLIA, M e CRESTI, E. L´intonazione e i criteri di trascrizione Del parlato adulto e infantile. In: BORTOLINI, U., PIZZUTO, E. *Il Progetto CHILDES Italia*. Pisa: Del Cerro, 1997, pp. 57-90.
MONEGLIA, M., SCARANO, A., SPINU, M. Validation by expert transcribers of the C-ORAL-ROM prosodic tagging criteria on Italian, Spanish and Portuguese corpora of spontaneous speech. In: *Information Society Technologies Programme (C-ORAL-ROM)*
RASO, T., MELLO, H., DE DEUS, L. e JESUS, A. Uma aplicação da Teoria da Língua em Ato ao português do Brasil. In: *Revista de Estudos da Linguagem*, 2007, pp.147-166.
RASO, T. e ULISSES, A. Tópico e Apêndice no português do Brasil: algumas considerações. In: *Revista de estudos da linguagem*. 2008, 247-262.
RASO, T. e MELLO, H. As especificidades da unidade de tópico em PB e possíveis efeitos do contato lingüístico. In: SARAIVA, E. e CHAVES MARINHO, J. *Estudos da língua em uso: da gramática ao texto*. In press.
ULISSES, A. J. *A unidade de Apêndice no português do Brasil*. Dissertação de Mestrado. UFMG, 2008 (Orientador: Tommaso Raso).
ZÁGARI, M. R. L. . Os Falares Mineiros: Esboço de Um Atlas Lingüístico de Minas Gerais. In: V. de ANDRADE AGUILERA. (ed.). A *Geolingüística no Brasil - trilhas seguidas, caminhos a percorrer*. Londrina: Editora da Universidade Estadual de Londrina, 2005, v. 1, p. 45-72.