C-ORAL-BRASIL I

Corpus de referência
do português brasileiro
falado informal

Tommaso Raso
Heliana Mello
Organizadores

C-ORAL-BRASIL I

Corpus de referência do português brasileiro falado informal

Belo Horizonte
Editora UFMG
2012

C787 C-ORAL-BRASIL I : *corpus* de referência do português brasileiro falado informal / Tommaso Raso, Heliana Mello, organizadores. – Belo Horizonte : Editora UFMG, 2012.
332 p. : il + 1 DVD-ROM. – (Linguajar)

Acompanha DVD com conteúdo completo do livro, arquivos de alinhamento, anotação gramatical, especificações do *corpus*, medidas estatísticas, listas de frequência e arquivos de som.
Inclui bibliografia.
ISBN: 978-85-7041-943-9

1. Linguística. 2. Comunicação oral. I. Raso, Tommaso. II. Mello, Heliana. III. Série.

CDD: 410
CDU: 81

Elaborada pela DITTI – Setor de Tratamento da Informação
Biblioteca Universitária da UFMG

Este livro recebeu apoio financeiro da Fapemig.

Coordenação editorial Danivia Wolff
Assistência editorial Eliane Sousa e Euclídia Macedo
Coordenação de textos Maria do Carmo Leite Ribeiro
Preparação de textos Alexandre Vasconcelos, Danivia Wolff e Maria do Rosário Alves Pereira
Revisão de provas Beatriz Trindade, Camila Figueiredo, Davi Bezerra de Souza
Coordenação, Projeto gráfico e Formatação Cássio Ribeiro
Capa e Produção gráfica Diêgo Oliveira

Editora UFMG
Av. Antônio Carlos, 6.627 – Ala direita da Biblioteca Central – térreo
Campus Pampulha – 31270-901 – Belo Horizonte/MG
Tel.: + 55 31 3409-4650 Fax: + 55 31 3409-4768
www.editora.ufmg.br editora@ufmg.br

Para o Nando
(*in memoriam*).

Sumário

CAPÍTULO 4 | *Heliana Mello | Tommaso Raso | Maryualê M. Mittmann Heloísa P. Vale | Priscila O. Côrtes*

Prefácio

É com grande satisfação que saudamos a publicação do *corpus* C-ORAL-BRASIL, fruto do trabalho intenso de Tommaso Raso e Heliana Mello e do entusiasmo do grupo de jovens do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da UFMG (Maryualê Mittmann, Heloísa Vale, Priscila Côrtes, entre outros).

Esta obra, que vimos começar em 2007-2008 e cujo desenvolvimento acompanhamos de perto dentro do acordo de colaboração entre a Università di Firenze e a UFMG, amplia, mesmo que de maneira independente, a coleção de línguas nacionais românicas da Europa (italiano, francês, português europeu, espanhol), representadas no C-ORAL-ROM, acrescentando a ela o português brasileiro (PB). Trata-se de um fato relevante para a fala românica em geral porque o C-ORAL-BRASIL permite uma representação mais definida do PB, à qual a comunidade científica poderá recorrer para marcar a distinção entre PB e português europeu (PE) – após séculos de contato do PB com línguas tão diferentes, em primeiro lugar ameríndias e depois africanas (cf. Mello, 1997 e 2011). É relevante também pela proeminência atual do Brasil na cena político-econômica internacional e pelo peso de 180 milhões de falantes, cujo uso linguístico pode ser retratado nacional e internacionalmente. Trata-se, enfim, de um fato relevante pelo nível de representação da fala alcançado nesta obra.

A escolha de desenvolver um novo recurso da fala espontânea brasileira seguindo o modelo do C-ORAL-ROM, em particular depois dos importantes trabalhos realizados no Brasil nesse domínio a partir dos anos de 1970 (como o NURC, entre outros), não era óbvia. Gostaríamos de lembrar aqui, rapidamente e do nosso ponto de vista, as motivações e o quadro cultural de referência que a bela descrição de Raso e o *background* ilustrado por Mello mostrarão em detalhe ao leitor mais à frente.

O *corpus* de fala românica C-ORAL-ROM é fruto de um projeto financiado pelo Quinto Programa Quadro da União Europeia, realizado por meio de um consórcio de entes públicos e privados (Università di Firenze, Universidad Autónoma de Madrid, Université Aix-Marseille en Provence, Centro de Linguística da Universidade de Lisboa, Fundação Bruno Kessler, Pitch France) concluído em 2004 com a publicação de um recurso multimídia que é distribuído pela indústria das línguas da European Language Resource Distribution Agency e na academia pela John Benjamins Publishing Company. O recurso permite o acesso simultâneo ao som e à transcrição de sessões gravadas das quatro coleções comparáveis de fala espontânea (italiano, espanhol, francês e português) e é acompanhado de instrumentos informáticos que permitem a análise acústica e a recuperação da informação textual para fins de estudo, aprendizagem e utilização para o desenvolvimento das tecnologias da linguagem.

A idealização do C-ORAL-ROM, no final dos anos de 1990, seguia uma série de relações científicas entre linguistas, em especial franceses, italianos e portugueses (junto conosco, Claire Blanche-Benveniste e Fernanda Bacelar do Nascimento, e, mais tarde, Francisco Marcos Marin), que reconheciam a urgência do estudo da realização falada da linguagem, a qual havia se tornado possível somente em tempos relativamente recentes, em função do desenvolvimento tecnológico. Esses estudiosos reconheceram, principalmente, o interesse teórico que o estudo apresentava, tanto para a descrição das línguas nacionais, quanto para os nossos conhecimentos gerais sobre a linguagem, tão diferente na sua realização oral em relação à sua idealização escrita, na qual sempre se basearam as gramáticas, e, a partir da segunda metade do século passado, a linguística da competência.

Essa exigência havia levado, nos 20 anos anteriores, à formação de vários recursos de fala, na Itália, em Portugal, na França e na Espanha, ligados a abordagens e tradições em parte diferentes, mas acomunadas pela perspectiva do estudo linguístico baseado em *corpora*. O C-ORAL-ROM deriva da necessidade de se colocarem à disposição dados mais estritamente comparáveis para o estudo da realização oral da linguagem em nível interlinguístico, sem os quais seria impossível evidenciar, com adequada base de observação, tanto as propriedades diferenciadoras da oralidade quanto os aspectos que, nessa forma de realização, diferenciam entre si as línguas naturais.

O C-ORAL-ROM fixou dois pontos nessa tarefa, definindo, por um lado, os requisitos nas modalidades necessárias para a representação do dado linguístico e, por outro, requisitos relativos à formação dos *corpora* para fins da sua comparabilidade. Ambos foram seguidos e tornados ainda mais eficazes na realização do C-ORAL-BRASIL, fazendo desse *corpus* um recurso plenamente compatível com seu antecedente. Vale a pena, aqui, ressaltar essas escolhas e a maneira exemplar como foram levadas adiante neste trabalho. Falaremos brevemente dos critérios de formação do *corpus* e com mais detalhe dos critérios de representação da fala. Ambos constituem a premissa para o desenvolvimento de estudos comparativos sobre a oralidade.

O requisito da comparabilidade na formação do *corpus* e, ao mesmo tempo, o requisito de representação da fala espontânea, juntos, constituem um desafio. Não é possível, de fato, constituir *corpora* paralelos e, portanto, forçadamente comparáveis, sem renunciar à característica principal da fala – a espontaneidade. A fala espontânea é caracterizada por grande variação, em conexão seja com as características do falante, seja com as características da situação em que acontece o ato comunicativo. Em particular, a fala se produz com base na interação, e a variedade dos atos de fala realizados na vida cotidiana e das suas situações eliciantes é ampla e desconhecida.

O C-ORAL-ROM respondeu a esses requisitos seguindo a regra de ouro da Linguística de *Corpus* para a formação de *corpora* de referência (veja-se o British National Corpus - BNC): definição dos critérios de variação utilizados para a representação do universo e assunção dos mesmos critérios na constituição de cada *corpus*.

As linhas ao longo das quais a fala varia são, contudo, muitas. A escolha dos critérios de amostragem é, portanto, um capítulo essencial para assegurar a representatividade dos recursos, seja em nível intralinguístico (representar a variação), seja em nível interlinguístico (representá-la da mesma maneira em cada língua).

Já o BNC, na sua seção de oralidade, individualizava dois critérios de amostragem a serem seguidos em paralelo: amostragem por contexto e amostragem por falante. De um lado, coletam-se amostras de fala de segmentos diastraticamente balanceados da população; de outro, identificam-se *a priori* contextos típicos nos quais a qualidade da produção oral é claramente diferente (bate-papos, fala oficial ou acadêmica, fala midiática,

entrevistas). Do cruzamento dos dois parâmetros, espera-se conseguir uma representação suficiente do universo.

Infelizmente, a prática mostrou que essa escolha, na realidade, reduz aos simples bate-papos o âmbito em que a variação da fala cotidiana é maior, ou seja, o uso informal da linguagem. A amostragem diastrática, de fato, produz amostras unicamente das situações mais fáceis de serem gravadas: os bate-papos em situação privada e familiar. Como a dificuldade do *corpus* israelense, o CoSIH (Izre´el; Hary; Rahav, 2001), mostrou, não se consegue, por muitos motivos que lembramos em várias ocasiões, gravar os falantes em todas as situações por eles encontradas durante o dia. Isto reduz enormemente a possibilidade de documentar, no *corpus*, a variedade dos contextos que, na vida cotidiana, determinam exigências interativas diferentes e produzem a exigência de atividades linguísticas as mais variadas. É essa propriamente a especificidade da fala espontânea.

Em outras palavras, na prática, a amostragem diastrática não produz variação diafásica, e, contrariamente a qualquer critério de amostragem, a variação diafásica nos contextos informais tem probabilidade zero de ocorrência, com a exceção dos bate-papos.

A escolha do C-ORAL-ROM e, com sucesso bem maior, a escolha do C-ORAL-BRASIL tentaram modificar o estado da arte dando espaço à maior variação diafásica possível. Por um lado, renunciou-se a um balanceamento diastrático da amostra, por outro, escolheram-se fixar critérios gerais que aumentassem as probabilidades de ocorrência do maior número possível de situações de interação específicas do uso informal da linguagem: a) variação por estrutura do evento comunicativo (diálogo, monólogo, multidiálogo); b) variação por contexto social em que o evento ocorre (vida pública *versus* vida privada e familiar). Cruzando os dois critérios e requerendo a maior variação possível de situações em cada ramificação da amostragem, conseguimos documentar o informal muito além dos bate-papos, com muitas amostras de fala fortemente interativa em situação (enquanto se conserta um carro, enquanto se briga, enquanto se dá aula de direção, enquanto se revelam fotografias, enquanto se tenta entrar em acordo com os amigos sobre o que comprar e como pagar, enquanto se lida com clientes em um bar, no guichê dos impostos, no trabalho etc.).

A documentação desse tipo de variação constitui o verdadeiro valor adjunto dos *corpora* de fala espontânea e, no estado atual, pode ser documentada em um *corpus*, e com muito trabalho, somente com base em

uma escolha precisa de amostragem. Os estudantes e os professores que os orientam, e é esse o caso do C-ORAL-BRASIL, são os vetores privilegiados que descobrem novas possibilidades de documentação, penetrando em seus ambientes de origem, necessariamente os mais variados em sociedades complexas. A casualidade requerida pelas amostragens desse tipo é limitada, portanto, pelas restrições do método, mas não tem como consequência reduzir a zero a probabilidade de ocorrência das situações de interação em que se usa a oralidade informal além dos bate-papos.

Acreditamos que o C-ORAL-BRASIL alcançou o maior resultado possível com essa metodologia, ultrapassando em qualidade e quantidade a variação diafásica dos *corpora* documentados no C-ORAL-ROM. Menos de 20% dos textos de interação dialógica ou conversacional no C-ORAL-BRASIL são constituídos por bate-papos ou entrevistas, enquanto grande parte das sessões de gravação documentam situações caracterizadas por acionalidade em movimento (veja-se o Capítulo 2), tornando possível até a gravação de um grupo de amigos enquanto jogam uma partida de futebol.

Dos muitos problemas que podem surgir em uma amostragem desse tipo, o C-ORAL-BRASIL evitou provavelmente o maior: documentar falantes pertencentes à mesma camada social, achatando a diversidade linguística que é função da diastratia.

Como se evidencia no Capítulo 2, o balanceamento diastrático, resultado do amplo número de falantes registrados no C-ORAL-BRASIL, mesmo que não perseguido intencionalmente, é um notabilíssimo íon. Ao contrário de como normalmente se pensa, na prática, não é tanto a variação diastrática que induz a variação diafásica, mas é esta última que induz a primeira.

Ressaltamos, então, não sem uma ponta de inveja, que o C-ORAL-BRASIL constitui um passo significativo no estado da arte da constituição dos *corpora* orais do ponto de vista dos resultados obtidos na amostragem do universo. Será preciso muito tempo para que a comunidade de estudiosos considere as consequências da disponibilidade desses dados para o estudo da linguagem e se adéque a esse "passo à frente".

O primeiro aspecto é de fácil compreensão. Não é possível estudar a oralidade perdendo a sua característica informação acústica ou sem uma representação capaz de fixar a volatilidade do evento fonético em unidades informacionais mínimas, as palavras. A solução não é, contudo, de realização fácil, porquanto necessita, por um lado, da adoção de formatos textuais

computáveis adequados à representação do diálogo e, por outro lado, de critérios de relação entre o evento fonético e a sua contraparte linguística. Não por acaso, a maior parte dos *corpora* de fala realizados nos anos de 1990 não permitiam o acesso ao sinal acústico, e mesmo a qualidade das transcrições era deficitária se comparada com a sua fonte.

Precisamos lembrar que a concepção dos *corpora* de fala mudou totalmente durante os últimos 20 anos. De fato, ainda no começo dos anos de 1990, a sua relevância era concebida essencialmente como possibilidade de acesso a recursos lexicais e, só parcialmente, sintáticos diferentes da escrita. Disto constituem prova os grandes *corpora* da época, dentre os quais o primeiro é o BNC, cuja estruturação se mantém ainda hoje exemplar, mas que, em sua seção de fala, que previa 10 milhões de *tokens* sobre os 100 do recurso total, não previa o acesso à contraparte acústica (disponível agora depois de duas décadas), e cujas transcrições, é um fato notório, foram feitas praticamente por funcionários e não por pesquisadores treinados. O mesmo pode ser dito do *corpus* alemão (Mannheim Corpus), dos anos de 1970, e, pelo que sabemos, do grande *corpus* brasileiro NURC. Todos estes apresentam, portanto, ausência total ou parcial do correspondente sonoro da gravação, que, se feito em suporte analógico, pode se desgastar completamente em uma década, além de transcrições ortográficas frequentemente parciais, realizadas com base em critérios não declarados, às vezes resumindo, às vezes normalizando, e raramente feitas com acurácia e em conformidade com o dado sonoro originário.

No C-ORAL-BRASIL, assim como no C-ORAL-ROM, a transcrição foi realizada segundo critérios ortográficos, ou seja, não fonéticos, e seguindo formatos padrão já a partir dos anos de 1980 (CHAT). A exploração de praxes consolidadas como o formato CHAT evita a introdução de convenções idiossincráticas na representação textual, que reduzem a possibilidade de utilização do recurso na comunidade científica e maximizam as possibilidades de computação automática da informação textual.

Vale então a pena lembrar, brevemente, algumas implementações desse formato que se confirmaram funcionais para a constituição de grandes *corpora* de fala românica. Primeiramente, a manutenção de uma sequência unitária para preservar o turno dialógico de cada falante, que deve ser entendido como a primeira entidade natural da fala; e em segundo lugar, a anotação sistemática, dentro da transcrição em nível zero, ou seja, no nível da transcrição da palavra, do que foi definido como *quebras prosódicas terminais*

e *não terminais* (//, /). As quebras indicam todas as rupturas perceptualmente relevantes para a avaliação da organização das produções faladas em agrupamentos finalizados do ponto de vista informacional.

A escolha da representação ortográfica, e não fonética, implica o desejo de estudar a oralidade com base em seu valor linguístico e não com base em sua primeira articulação. Nisto, o C-ORAL-BRASIL se encontrou em um ponto de fronteira difícil, e o valor e o interesse do que foi proposto devem ser ressaltados. Se a transcrição fonética torna obscuros os elementos da cadeia da fala aos quais atribui valor linguístico, seguir a ortografia padrão definida com base em uma idealização abstrata da língua tornaria inutilizável o dado vivo da fala para evidenciar os processos morfossintáticos próprios do português brasileiro, obscurecendo processos dinâmicos que caracterizam componentes inteiros dessa variedade, a morfologia livre e flexiva em primeiro lugar.

A transcrição da fala de línguas como o italiano, o francês, o espanhol e o PE pôde se servir de convenções de transcrição, mesmo se não compartilhadas unanimemente, de variedades regionais afirmadas e consolidadas e, de toda maneira, não muito distantes do padrão nacional. Não era esse, contudo, o caso do *corpus* de fala brasileiro. A questão envolve vários níveis: em primeiro lugar, a coleção brasileira é de fala mineira, mais especificamente, da cidade de Belo Horizonte e de sua área metropolitana – tal como requisitado por cada *corpus* de fala, que deve ser representativo de uma única variedade regional. Parece que muitas tendências gerais em curso na fala brasileira foram primeiro atestadas exatamente nessa variedade regional. Portanto, nos encontraríamos frente a uma situação de "vanguarda" linguística do mineiro que, evidentemente, deve ser registrada. Entretanto, não é fácil distinguir entre o que é estável dentro de uma variedade e, portanto, relativo ao sistema, e o que pode ser uma variante idiossincrática de um falante, que, ao contrário, deve ser normalizada, para não comprometer a reconhecibilidade.

Entretanto, o problema mais importante é que a fala mineira, como uma das variedades mais significativas do PB, é muito distante da de Lisboa ou *tout court* do PE. Não compete a nós, neste Prefácio, enfrentar a questão de qual é a relação entre PB e PE. Certamente, tivemos a possibilidade de verificar pessoalmente a distância entre um e outro. Uma transcrição da fala mineira, por mais fiel que seja ao conteúdo linguístico de entidades terminadas alinhadas com o som, impõe o problema de o quê e como transliterar.

O sistema pronominal sujeito nas suas variantes tônica e átona, a falta da marca de plural em nomes e adjetivos, a redução da morfologia verbal, a nova declinação verbal para alguns lemas ou tempos, o léxico diferente, mas também palavras existentes em PE cuja forma não é mais a mesma, fórmulas, interjeições, alocutivos autóctones, ou seja, a condição de ter que refletir o som por um lado e o significado por outro induziu os autores a se confrontarem com um novo sistema complexivo de escrita ortográfica, tarefa de peso significativo. Sentimos vontade de dizer, a esse respeito, que as soluções escolhidas nos parecem muito conscientes e compartilháveis, além de testemunharem coragem e cautela ao mesmo tempo. O resultado é uma transcrição correspondente ao som originário, legível e compreensível para qualquer falante brasileiro, que, contudo, achará traços específicos da variedade mineira que podem não ser os seus. Provavelmente a transcrição parecerá também escandalosa a quem se atém às convenções do português escrito padrão, e podemos imaginar que não faltará quem diga que os textos assim transcritos são obscuros e incompreensíveis. Já aconteceu, para os textos italianos falados na variedade toscana, que a presença de algum apóstrofo para sinalizar aférese e truncamento, algumas poucas variantes na morfologia verbal, algum lema regional, locução verbal ou fraseologia não pan-italiana, mas absolutamente transparente, levasse o *corpus* a ser considerado como dialetal e sem validade representativa.

As convenções, nos casos em que se decidiu por uma representação ortográfica não padrão da palavra no C-ORAL-BRASIL, ilustradas no Capítulo 4, foram fruto de reflexão intensa por parte dos organizadores (veja-se também Mello e Raso, 2009). Elas são, ao mesmo tempo, garantia da qualidade das transcrições e uma premissa para a utilização dos dados do *corpus* oral para uma descrição gramatical. Voltaremos a esse último ponto, mas, quanto às garantias sobre a confiabilidade da transcrição, o C-ORAL-BRASIL apresenta um resultado absolutamente original no domínio dos *corpora* orais. A adequação da transcrição foi testada e é garantida não somente, como normalmente é feito, com relação à sua computabilidade (vejam-se os resultados relativos à etiquetagem morfossintática no Capítulo 6), mas também frente ao dado acústico. Os resultados dos testes de validação, tanto gerais quanto nas formas não padrão, evidenciam um nível de adequação além de qualquer requisito: uma margem de erro geral inferior a 0,81% e inferior a 0,57% para as formas relativas aos critérios não padrão implementados pelos autores,

sendo que o fenômeno que apresentou a maior porcentagem de erro (as preposições apostrofadas) não superou os 3,28% de erros (veja-se o Capítulo 4). Claire Blanche-Benveniste teria gostado desse nível de acurácia, ela que foi a primeira a transferir uma tradição filológica de precisão e transparência nos critérios de transliteração. Os linguistas que desejarem utilizar o *corpus* C-ORAL-BRASIL para identificar as mais consistentes mudanças do PB com relação ao PE poderão, portanto, confiar nos dados validados quanto à uniformidade ortográfica dos traços morfológicos característicos.

A nossa avaliação é que esse trabalho marca um limite a partir do qual existirá um *ante quem* e um *post quem*, a partir do qual se tornará mais fácil realizar também novas coleções de diferentes variedades brasileiras e conduzir pesquisas gramaticais que deverão levar em conta os dados linguísticos, que terão rapidamente mudado em uma sociedade em expansão como a brasileira. A impressão é de que, devido à influência que contatos com línguas, principalmente africanas, geraram no português e seu aprendizado como segunda língua e não como língua materna pela maior parte da população, o atual uso pervasivo do português deva ser associado a uma inevitável e profunda transformação, de que o *corpus* de fala mineira constitui um testemunho precioso.

Apesar de a escolha da transcrição ortográfica denunciar o interesse desse *corpus* pelo estudo linguístico, os foneticistas e os tecnólogos que queiram utilizá-lo para o estudo da voz não ficarão decepcionados e poderão usufruir da altíssima qualidade da fonte acústica. A dedicação e a acurácia com as quais foram efetuadas as gravações possibilitaram um resultado até pouco tempo inimaginável para um grande *corpus* de fala espontânea, que é obrigado a gravar uma ampla variedade de situações diafásicas e contextuais, cada uma necessitando de soluções específicas de microfonação. O uso de microfones monodirecionais, praticamente um para cada falante, eliminou em grande parte os fenômenos de retorno em todos os textos, sempre identificados com um metadado relativo à qualidade acústica. Tivemos a possibilidade de ter acesso ao dado acústico de várias sessões de gravação incluídas no C-ORAL-BRASIL, e os espectrogramas quase não se diferenciam dos traçados obtidos com base em gravações em câmara anecoica. Por isso, o recurso resulta adequado à transcrição fonética e ao alinhamento silábico para os estudiosos que desejarem utilizá-lo com esse objetivo no futuro.

Na representação da fala, o alinhamento do som à transcrição é o ponto, teoricamente, mais delicado, sendo realizado no C-ORAL-ROM e no C-ORAL-BRASIL segundo um critério e uma técnica comuns.

O desenvolvimento vertiginoso das possibilidades técnicas permitiu transformar essa exigência naquilo que hoje nos parece um requisito correto para um bom usufruto e uma boa apreciação do texto oral. A técnica utilizada é fornecida pelo *software* WinPitch, de Philippe Martin, adaptado à função de alinhamento de grandes *corpora* para o projeto C-ORAL-ROM. O WinPitch permite o alinhamento em níveis independentes de porções do sinal e oferece a função de sincronia som-transcrição, acompanhada pela análise acústica do sinal em tempo real.

O alinhamento produz dois efeitos fundamentais e que, em parte, poderiam parecer contraditórios. Por um lado, a presença do som obriga a uma transcrição em conformidade com a modalidade original. Por outro, o fato de que uma transcrição de fala espontânea quase nunca pode ser considerada definitiva, porque a escuta e a reescuta induzem, quase sempre, a pequenas correções, leva à consciência de que ela não é inteiramente autossuficiente e deve sempre ser acompanhada pelo som. A qualidade da transcrição no C-ORAL-BRASIL depende, também, das revisões realizadas após o alinhamento.

O alinhamento implica também o problema de o quê é alinhado ao quê. A escolha do enunciado como unidade de alinhamento corresponde à escolha desse nível pragmático de descrição linguística como unidade de referência da língua falada (Cresti, 2000a e 2000b). Como bem argumenta Raso, no Capítulo 3, a Teoria da Língua em Ato oferece uma solução operacional para a identificação do enunciado, através da identificação das unidades concluídas por quebras prosódicas terminais (enunciados, padrões ilocucionários, estrofes). Estas constituem as unidades de referência alinhadas ao som: de silêncio ou quebra prosódica terminal ao sucessivo silêncio ou quebra prosódica terminal.

O critério de definição dos enunciados no C-ORAL-BRASIL, como no C-ORAL-ROM, é, portanto, prosódico e se fundamenta na percepção das rupturas entonacionais da cadeia falada percebidas como terminais pelo transcritores.

A introdução, na representação falada, dos critérios perceptuais, especificamente em pontos cruciais como a identificação das unidades de referência, essenciais para qualquer análise linguística superior ao nível lexical, frequentemente deixava perplexos muitos colegas, mesmo os

mais prestigiosos, que, por um lado, temiam a introdução de fatores idiossincráticos e não verificáveis e, por outro, estavam acostumados a utilizar para esses fins índices sintáticos (Biber) ou interpretativos (Blanche-Benveniste) sentidos como "mais linguísticos".

Notar a correlação estrita entre unidades prosódicas e atividades pragmáticas é o ponto essencial que caracteriza a série de recursos C-ORAL-BRASIL / C-ORAL-ROM e corresponde à concepção mais geral da fala espontânea como interação pragmática, mais que como atividade de construção textual.

A correlação entre entidades linguísticas dotadas de força ilocucionária e, portanto, de autonomia pragmática e unidades linguísticas concluídas por uma quebra prosódica percebível é um dado de fato. O WinPitch permitirá aos usuários verificar por conta própria a adequação da segmentação do fluxo da fala em enunciados proposta no C-ORAL-BRASIL, mas o recurso é acompanhado por uma série de medidas de validação e dos resultados em termos de acordo entre os anotadores, que deixam poucas dúvidas a respeito da sua confiabilidade. O C-ORAL-BRASIL dedicou uma parte importante das atividades de constituição do *corpus* à validação e ao refinamento dos resultados obtidos quanto à confiabilidade da partição do dado acústico em unidades discretas com base prosódica (veja-se o Capítulo 4 deste volume).

A proeminência das quebras prosódicas em geral foi verificada em muitas ocasiões. Quanto às pesquisas relativas à constituição dos grandes *corpora* de fala, o DUTCH *corpus*, no começo da década passada, apresentou dados relevantes com relação à detecção das quebras por parte de não *experts*. Especificamente a proeminência das quebras terminais para a percepção e, em medida consistente, também a proeminência das quebras não terminais, foram verificadas, sempre com não *experts*, nas coleções C-ORAL-ROM. Também foram replicadas como um dado relevante interlinguisticamente para além do quadro românico (vejam-se os trabalhos de Izre'el sobre o hebraico moderno (CoSIH) e a recente constituição do *corpus* das línguas afro-asiáticas (CorpAfroAs)). O trabalho desenvolvido para o PB não nasce, portanto, do nada, mas a estratégia de anotação e revisão adotada nesse caso é original e permitiu alcançar resultados qualitativamente muito superiores ao modelo do qual descende.

A percepção das quebras, de fato, não corresponde à percepção de um dado positivo, como, por exemplo, seria uma real interrupção do fluxo da

fala (pausa), mas a uma constelação de qualidades. *Reset* prosódico, queda de intensidade, proeminências prosódicas adjacentes e alongamento silábico são propriedades prosódicas, mas o cumprimento do ato de fala e a presença de fronteiras sintáticas são também correlatos que acompanham a percepção de uma quebra. Pode, portanto, ser difícil para um transcritor distinguir a informação prosódica relevante de outras informações, e isso pode levar tanto a fenômenos de sobre-extensão das marcas prosódicas na transcrição quanto à sua subextensão e, assim, resultar em uma diminuição da coerência das anotações. Raso e Mittmann (2009 e no Capítulo 4 deste livro) mostraram que essa coerência aumenta muito se os grupos de anotadores se comportam de maneira experiente e adequam seu horizonte perceptual, alcançando entre os anotadores um acordo nas fases anteriores à elaboração do *corpus*. Esta praxe deverá ser levada em conta na compilação de recursos futuros, e, em particular, a formação dos transcritores deverá ser considerada como um requisito essencial para a formação de *corpora* orais, se se quiser que os *corpora* se adéquem ao padrão de qualidade do C-ORAL-BRASIL.

A transcrição do C-ORAL-BRASIL contém sinais diacríticos de ruptura prosódica terminal e não terminal. O sistema é o mesmo aplicado a todos os *corpora* europeus do C-ORAL-ROM. Este requisito é central, a nosso ver, para o estudo da estrutura primária que caracteriza a organização da fala antes da sintaxe, ou seja, a articulação da informação. A escansão prosódica, que, repetimos, é um dado perceptual imediato e proeminente, é considerada na Teoria da Língua em Ato como a marca necessária da organização informacional com base pragmática dos enunciados e de entidades mais extensas da textualidade falada (padrões ilocucionários e estrofes). Sinteticamente, a arquitetura geral pode ser assim esquematizada:

- ato perlocucionário: ativação afetiva (pulsão) do ato da linguagem como um todo;
- ato ilocucionário: através do impulso afetivo, a ativação de um esquema acional segundo tipologias ilocucionárias convencionalizadas e a organização de um padrão informacional composto por várias unidades informacionais com diferentes funcionalidades pragmáticas, cujo núcleo deve ser constituído por uma unidade necessária e suficiente dedicada ao cumprimento da força ilocucionária do enunciado;

- interface prosódica obrigatória entre a sinalização do cumprimento do ato ilocucionário, as diversas unidades informacionais que podem compor o ato, e a sua execução locutiva;
- ato locutivo: através da sinalização prosódica, ativam-se entidades semântico/sintáticas correspondentes às unidades de informação componentes do padrão informacional e cumpre-se o ato ilocucionário, ou as outras unidades textuais maiores.

No Capítulo 3, Raso ilustra em detalhe as premissas teóricas e se aprofunda na descrição das diferentes funções informacionais, com uma rica exemplificação retirada do *corpus*. As funções informacionais podem ser classificadas segundo duas tipologias primárias: aquelas de tipo textual (Comentário, Tópico, Apêndice de Comentário e de Tópico, Parentético, Introdutor Locutivo) e aquelas de suporte dialógico (Incipitário, Fático, Conativo, Alocutivo, Expressivo, Conector Discursivo), que foram identificadas ao longo dos anos de trabalho experimental conduzido sobre o *corpus* italiano e sobre amostras espanholas e francesas, mas que, a esta altura, acharam confirmação na análise sistemática conduzida sobre o *corpus* de fala do PB. O C-ORAL-BRASIL dará, e já está dando, uma contribuição essencial ao estudo interlinguístico da articulação da informação na fala. Com base no C-ORAL-BRASIL e no C-ORAL-ROM italiano, foi possível constituir duas bases de dados comparáveis de fala informal de 5 mil enunciados por *corpus* etiquetados informacionalmente. No VII International LABLITA Workshop, realizado em Florença, em junho de 2011, foram apresentados os primeiros trabalhos de caráter geral sobre a estrutura informacional de línguas românicas como o italiano, o PB e o francês. Os colegas da equipe mineira começaram então a evidenciar peculiaridades da organização informacional, mas, consequentemente, também semântico-sintática, da variedade brasileira com relação às outras línguas românicas. O trabalho está só no começo, mas se anuncia como um dos pontos de maior novidade e relevância no estudo da fala.

Emanuela Cresti
Massimo Moneglia
(LABLITA, Università di Firenze)