O projeto C-ORAL-BRASIL

Tommaso Raso°, Heliana Mello°, Maryualê Mittmann*
°UFMG/CNPq/Fapemig, *UFMG

This paper reports on the C-ORAL-BRASIL Project, a spontaneous speech corpus compilation effort to document Brazilian Portuguese. C-ORAL-BRASIL is a sister project to the European C-ORAL-ROM, and follows its same architecture and compilation parameters. In this paper C-ORAL-BRASIL features are described, the major ensuing studies from it are presented and the new research and corpus compilation efforts under way are introduced.

Keywords: spontaneous speech, oral corpus, Brazilian Portuguese, C-ORAL-BRASIL

1. Introdução

O projeto C-ORAL-BRASIL é coordenado por Heliana Mello e Tommaso Raso, e está sediado no Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL), na Faculdade de Letras da UFMG. O objetivo geral do projeto é o estudo da fala espontânea através da compilação e análise de corpora.

Desde 2008, o projeto está inserido em um acordo institucional entre a Universidade Federal de Minas Gerais e a Università degli Studi di Firenze coordenado por Emanuela Cresti e Tommaso Raso. De fato, o C-ORAL-BRASIL nasceu como projeto irmão do C-ORAL-ROM (Cresti & Moneglia 2005) e dos projetos desenvolvidos no Laboratório LABLITA[1]. Desde então, os dois grupos têm promovido trocas constantes e têm estudado a fala espontânea com as mesmas metodologias, os mesmos objetivos científicos e o mesmo arcabouço teórico (Cresti 2000; Moneglia 2011; Moneglia & Raso in press).

Além das trocas de visitas mútuas para o desenvolvimento das atividades de pesquisa, os dois grupos organizaram vários eventos, com a participação também de outros grupos. Entre aqueles organizados no Brasil citamos o V

[1] http://lablita.dit.unifi.it/

CHIMERA. Romance Corpora and Linguistic Studies 1 (2014), 31-67. ISSN 2386-2629

LABLITA Workshop em 2010[2], o 7th International GSCP Conference em 2012[3], e o 3rd Methological LEEL Seminar em 2013. O LEEL também organiza periodicamente cursos para aperfeiçoamento dos seus próprios membros em disciplinas normalmente não oferecidas pelos canais institucionais, principalmente disciplinas de estatística aplicada aos estudos linguísticos e de fonética.

Neste artigo apresentaremos os principais resultados do projeto obtidos até o final de 2013, começando pela realização do corpus C-ORAL-BRASIL (Raso & Mello 2012) e passando depois aos estudos sobre ele (e sobre outros corpora da família C-ORAL). Falaremos com mais detalhes dos estudos sobre a estrutura informacional, sobre as ilocuções e as atitudes, e sobre a modalidade, mas trataremos também de outros estudos que estão sendo realizados. O objetivo é mostrar como a compilação de um corpus com os parâmetros do C-ORAL-BRASIL permite realizar novos estudos, impossíveis sem um corpus com essas características, ou verificar, utilizando metodologias novas, os resultados de estudos sobre temas muito frequentados na literatura.

2. O corpus C-ORAL-BRASIL

O C-ORAL-BRASIL (Raso & Mello 2012) é um corpus de fala espontânea do português brasileiro, representativo da diatopia mineira, sobretudo da região metropolitana da cidade de Belo Horizonte, capital do estado de Minas Gerais[4]. A compilação deste corpus seguiu os mesmos parâmetros adotados pelo projeto C-ORAL-ROM para corpora de espanhol, francês, italiano e português europeu (Cresti & Moneglia 2005), sendo desta forma comparável em arquitetura e amostragem a esses corpora. Portanto, o C-ORAL-BRASIL se configura como um corpus de terceira geração, com o texto e o áudio alinhados.

O C-ORAL-BRASIL se presta a estudos linguísticos de natureza variada, entretanto é especialmente interessante para o estudo de aspectos relacionados à pragmática e estrutura informacional por se tratar de um corpus de fala

[2] O produto desse workshop foi o volume *Pragmatics and Prosody. Illocution, Modality, Attitude, Information Patterning and Speech Annotation*, organizado por Mello, Panunzi e Raso e publicado em 2012.

[3] Produtos desse congresso foram o volume das Atas (Mello, Pettorino & Raso 2012) e um importante volume no prelo pela John Benjamins (Raso & Mello in press).

[4] O projeto C-ORAL-BRASIL foi financiado pela Fapemig, CNPq e UFMG.

espontânea de grande diversidade diafásica, segmentado em unidades tonais e enunciados.

A parte informal do C-ORAL-BRASIL, já publicada (Raso & Mello 2012), é composta pelos seguintes elementos[5]:

1. corpus multimídia (arquivos de som formato wav, arquivos de transcrição formatos rtf e txt, arquivos de alinhamento som-transcrição formato xml gerados pelo software WinPitch (Martin 2013));
2. metadados de cada arquivo de som (título, nome do arquivo, participantes e seus dados – gênero, idade, escolaridade, profissão e papel desempenhado na interação documentada - data de gravação, local, contexto e tópico, duração, número de palavras, qualidade acústica, nomes de transcritores e revisores, comentários sobre as peculiaridades da gravação;
3. transcrições do corpus anotadas morfossintaticamente através do parser PALAVRAS (Bick 2012) em formato xml e txt;
4. listas de frequência, medidas e estatísticas dos informantes;
5. relatório contendo as explicações técnicas para todos os parâmetros adotados na compilação do corpus;
6. um livro em formato pdf com a descrição do corpus, apresentação da teoria na qual está ancorado, explicações sobre os critérios de transcrição e de segmentação além de sua validação, discussão das principais medidas de fala obtidas no corpus e descrição do parser utilizado para a etiquetagem morfossintática.

As transcrições foram feitas seguindo-se o modelo CHAT (MacWhinney 2000), implementado para a anotação prosódica (Moneglia & Cresti 1997). O corpus é segmentado em unidades tonais e unidades terminadas (Raso 2012). A fronteira entre unidades tonais, correspondente à percepção de uma quebra prosódica com valor não conclusivo, é marcada por uma barra simples (/), enquanto a fronteira entre unidades terminadas, correspondente à percepção de uma quebra prosódica com valor conclusivo, é marcada por barras duplas (//). As unidades terminadas podem ser de dois tipos: enunciados e Stanzas.

[5] O C-ORAL-BRASIL está disponível para consulta online nas plataformas CorpusEye e Linguateca. Um subcorpus balanceado do C-ORAL-BRASIL anotado informacionalmente e denominado Minicorpus, está disponível através da plataforma DB-IPIC. Em breve, o corpus estará disponível também via TalkBank de Brian MacWhinney e em uma plataforma do projeto C-ORAL-BRASIL.

O conceito de enunciado aqui adotado pode ser definido como a menor unidade da fala interpretável pragmaticamente em autonomia (Cresti 2000; Moneglia & Martin 2005; Cresti & Gramigni 2004; Raso 2013), e corresponde portanto a um ato de fala (uma ilocução ou padrão ilocucionário). Essa definição é independente da sua configuração semântica e sintática. Na fala espontânea, principalmente nos monólogos, onde a interatividade se reduz sensivelmente, a unidade terminada, por assim dizer, se dilata no que é chamado de Stanza (Cresti 2010), sequências de mais de um ato de fala ligados por sinal de continuidade na mesma unidade terminada. Os exemplos da Tabela 1 podem esclarecer melhor os conceitos de enunciado e de Stanza.

2.1 A arquitetura do C-ORAL-BRASIL

Sendo um corpus de fala espontânea, o C-ORAL-BRASIL documenta fala planejada no momento de sua produção, i.e., fala que não é desempenhada a partir de texto parcial ou integralmente preparado, como seria o caso de textos performáticos ou discursos previamente desenvolvidos (Nencioni 1983; Cresti 2000; Biber 1988; Blanche-Benveniste *et al.* 1990; Miller & Weinert 1998; Givón 1979; Moneglia & Martin 2005; Moneglia 2011). Os eventos de fala que podem ser considerados espontâneos contêm:

1. interação multimodal face-a-face;
2. referência intersubjetiva ao espaço dêitico;
3. programação mental síncrona ao desempenho vocal;
4. comportamento linguístico contextualmente subdeterminado, i.e., comportamento imprevisível.

Uma longa tradição de estudos sociolinguísticos (Berruto 1987; Biber & Conrad 2001; Biber *et al.* 1998; Gadet 1996a, 1996b, 1997, 2000, 2003; Halliday 1989) debruçou-se sobre a parametrização para a caracterização de variedades de fala. Para a fala espontânea são apontados os seguintes parâmetros:

a. possibilidades estruturais dos eventos de fala (monólogos, diálogos, conversações);
b. canal comunicativo;
c. contexto sociológico, i.e., o domínio social do evento (familiar, privado, público);

d. condições de programação (fala parcialmente e totalmente programada , ou fala não-programada);
e. variedades de registro e gênero;
f. fatores sociolinguísticos caracterizadores dos falantes (gênero, idade, escolarização, ocupação);
g. origem geográfica;
h. objetivo do evento de fala;
i. assunto.

O planejamento de um corpus de fala espontânea é, como se vê, uma tarefa complexa que inclui a necessidade de se registrarem as principais variações presentes nos diferentes tipos de eventos de fala (Berruto 1987; Biber 1988; De Mauro *et al.* 1993; Gadet 1996a, 1996b, 2003).

Diferentemente dos corpora de fala produzidos em condições ambientais controladas que garantem alta qualidade acústica, como é o caso daqueles que documentam apenas interações telefônicas, map tasks, ou entrevistas controladas, corpora de fala espontânea são compilados a partir de dados obtidos em contexto natural, o que necessariamente reduz a qualidade acústica das gravações e apresenta outros desafios de natureza diversa, como interrupções, surgimento de interagentes inesperados, dentre outros. O C-ORAL-BRASIL foi produzido através de gravações em que foram utilizados equipamentos de última geração e microfones de lapela sem fio para que se obtivesse a melhor qualidade acústica possível, dadas as condições naturais de gravação.

Um dos objetivos importantes do C-ORAL-BRASIL é o de se conseguir comparabilidade com os corpora do projeto C-ORAL-ROM. Porém, assumindo-se que a documentação da fala espontânea necessariamente busca a máxima variação textual possível, consequentemente quanto maior for a diversidade textual, menor a comparabilidade entre textos. Assim, a comparabilidade entre corpora de fala espontânea se dá através da manutenção dos seus parâmetros de compilação e não através da similaridade textual, como é o caso de corpora escritos comparativos. Desta forma, o princípio seguido pelo C-ORAL-BRASIL foi o de similaridade de parâmetros compilatórios em relação aos corpora C-ORAL-ROM.

Até o momento, somente a parte informal do C-ORAL-BRASIL foi publicada, entretanto a sua parte formal já está sendo compilada. O corpus informal é composto por 208.130 palavras, distribuídas em 139 textos de aproximadamente 1.500 palavras cada um, em média. Alguns poucos textos são maiores (chegando até a 5.000 palavras) ou menores (porém mantendo a sua

autonomia textual). Os 139 textos são divididos em dois contextos: privado/familiar (159.364 palavras) e público (48.766 palavras); para cada um dos contextos os textos foram divididos igualmente entre as três tipologias interacionais: monólogos, diálogos e conversações (textos dialógicos com mais de dois interagentes).

Os textos são transcritos utilizando-se o formato CHILDES (MacWhinney 2000) implementado para a anotação prosódica (Moneglia & Cresti 1997). A anotação prosódica, além das marcações de unidades terminadas e de unidades tonais não terminais, apresenta a marcação de sequências interrompidas (+) e das retrações ([/n], onde n representa o número de palavras apagadas pelo falante durante a retração). A segmentação foi submetida a teste de validação Kappa, alcançando um valor de acordo de 0,86 (Raso & Mittmann 2009; Mello *et al.* 2012).

As transcrições seguem a ortografia tradicional do português brasileiro, com exceções significativas que tentam capturar fenômenos de gramaticalização e lexicalização em andamento nesta variedade linguística (Mello & Raso 2009; Mello *et al.* 2012).

2.2 A variação diafásica no C-ORAL-BRASIL

Um corpus de fala espontânea deve retratar a variação situacional o máximo possível. Isso se deve ao fato de que a estrutura da fala não é condicionada prioritariamente pela variação entre os falantes e nem mesmo pelo assunto de que se fala. Sobretudo, sob uma perspectiva pragmática, é crucial que se documentem as diferenças de comportamento verbal que dependem das variadas ações verbais executadas pelos falantes em situações distintas.

Enquanto a tradição sociolinguística nos permite classificar os principais domínios da fala formal, quando se trata de fala informal, as categorias permanecem abertas, não sendo possível listarem-nas. Assim, na compilação de um corpus de fala espontânea informal, documenta-se um amplo espectro de variação porque não se pode considerar um dado contexto como sendo mais típico que outro. Para que isso seja possível, dado o custo de realização de um corpus falado com relação a corpora escritos (não só na perspectiva econômica, mas principalmente levando-se em conta o tempo necessitado para as gravações, as transcrições e os alinhamentos), é importante apresentarem-se muitos textos diferentes e reduzir o seu tamanho. A medida média de 1.500 palavras é suficiente para que uma interação seja textualmente autônoma, e ao mesmo tempo permite que seja representada uma ampla variedade situacional. Em

outras palavras, quanto mais ampla é a variedade dos textos, mais apurada é a representação do universo textual possível; mas os textos devem também possuir uma dimensão suficiente para garantir que neles estejam presentes as características de um determinado contexto que se pretenda representar com eles. Um texto deve, ao mesmo tempo, poder apresentar determinadas propriedades sintáticas (micro-sintaxe) e determinadas estruturas dialógicas (macro-sintaxe e estrutura informacional) (Blanche-Benveniste *et al.* 1990; Scarano 2003).

Dentro do registro informal, a divisão entre contexto familiar/particular e contexto público documenta o papel com o qual um falante interage com outros: se ele interage como indivíduo, como acontece nas relações familiares ou com os amigos, ou se ele interage com base em um papel profissional ou institucional, como acontece, por exemplo, exercendo o papel de vendedor ou cliente, de cidadão e funcionário público, de aluno e professor, de garçom e cliente, de colega de trabalho etc. Considerando-se que na maior parte do tempo e na maior variedade de situações comunicativas apresentam-se interações de caráter familiar/particular, cerca de 70% do C-ORAL-BRASIL foram dedicados a esse contexto, deixando-se uma porcentagem menor aos textos do contexto público.

2.3 Variabilidade Textual

Observam-se a seguir, a partir da variabilidade textual presente no C-ORAL-BRASIL, aspectos qualitativos, com relação, principalmente, à variação diafásica e diastrática, e, em menor medida, à variação diatópica representadas neste corpus.

Como já mencionado, o C-ORAL-BRASIL procura representar ao máximo a variação de atividades efetuadas através da fala. Assim buscou-se reduzir ao máximo possível, principalmente para os diálogos, a incidência de bate-papos e de entrevistas, ou seja, situações em que os participantes da interação não cumprem outra atividade além da de falar; essas situações constituem a situação mais documentada nos corpora orais em geral, por ser a mais fácil de ser gravada; entretanto ela é também a menos interessante em uma perspectiva diafásica e pragmática.

Dentre os diálogos presentes no C-ORAL-BRASIL, somente 8, de um total de 48 textos, podem ser considerados da tipologia bate-papo ou entrevista. Nas conversações, a incidência de bate-papos é maior, porque essa tipologia, contrariamente à dialógica, caracteriza-se por ser mais frequentemente uma

tipologia sem ação específica. Para essa tipologia, 17 textos dentre 42 podem ser considerados bate-papos. Observe-se que nenhuma das conversações públicas pode ser considerada bate-papo.

Uma atenção específica foi dada para a gravação de interações em movimento, já que se deve considerar que a acionalidade estática e a acionalidade dinâmica constituem macro-domínios acionais diferentes: 4 diálogos foram gravados totalmente ou parcialmente dentro de carros em movimento; no total foram 19 as gravações em diferentes situações de movimento - entre as conversações há as seguintes gravações: um grupo de amigos jogando sinuca, um grupo de amigos jogando futebol, um grupo de pessoas preparando um almoço, uma visita ao centro de doação de sangue, uma compra em uma farmácia e uma aula de ginástica. Entre os diálogos, além daqueles já indicados, há as seguintes gravações: duas amigas fazendo compras em um supermercado, duas empregadas domésticas limpando a cozinha e outras partes da casa, um corretor que leva a irmã para visitar vários apartamentos, mãe e filha limpando seu apartamento, compras de sapatos e roupas em uma loja, uma aula de ginástica com um personal trainer, visita a um apiário e dois garçons preparando pizzas em uma festa e as oferecendo aos convidados.

2.4 O minicorpus e a etiquetagem informacional

Para a pesquisa sobre a estrutura informacional e as ilocuções na fala espontânea, foi realizada uma etiquetagem informacional de um minicorpus de 20 textos do C-ORAL-BRASIL, que refletisse o máximo possível o balanceamento do corpus. Em total, o minicorpus é composto por cerca de 30.000 palavras e 5.000 estruturas terminadas.

O minicorpus brasileiro é comparável a um minicorpus do italiano (para a compilação de ambos, veja-se apresentação do database IPIC por Panunzi & Gregori 2011; Gregori & Panunzi 2012; Panunzi & Mittmann in press), e agora também a um minicorpus de inglês, cuja etiquetagem está sendo completada. Os minicorpora constituem o principal recurso para o estudo das unidades informacionais (veja-se 3) e de outros estudos. O processo de etiquetagem com base na Language into Act Theory (L-AcT; Cresti 2000; Moneglia 2005, 2011; Moneglia & Raso in press) foi realizado manualmente pelas equipes do LEEL e do LABLITA.

3. O estudo das Unidades Informacionais

Sob a perspectiva de L-AcT, o estudo da estrutura informacional está ligado à identificação da segmentação prosódica do fluxo da fala. Em princípio, a cada unidade tonal corresponde uma unidade informacional, caracterizada por sua função dentro do enunciado (ou da unidade terminada), sua distribuição e seu perfil prosódico. Assim, a análise de parâmetros prosódicos é um passo fundamental para o estudo das unidades informacionais. Tal análise é possibilitada pelo alinhamento do som com a transcrição da fala, realizado, no C-ORAL-BRASIL, através do programa *WinPitch* (Martin 2013). A análise acústica, propriamente, é também realizada utilizando-se o programa *Praat*. Os procedimentos adotados para o estudo das unidades informacionais, de modo geral, envolvem a recuperação dos enunciados que apresentam a unidade em estudo no WinPitch e a exportação do som dos enunciados para o Praat. É realizada então a segmentação silábica da unidade e procede-se à medição de diversos parâmetros acústicos, quais sejam: duração de cada sílaba e total da unidade, intensidade (em dB) de cada sílaba, frequência fundamental (F0, em Hz e/ou em semitons) média, inicial e final de cada sílaba e de toda a unidade. Conforme a necessidade, também se realiza a estilização da curva da F0 e testes de aceitabilidade perceptual variando-se algum dos parâmetros.

3.1 Unidades que expressam força ilocucionária

As unidades informacionais cuja função é expressar a força ilocucionária do enunciado são consideradas o núcleo de um enunciado (veja 4).

A Tabela 1 apresenta exemplos[6], em PB, das unidades nucleares de enunciados – COM e CMM – e de *Stanzas* – COB. Os três conceitos (que podem ser aprofundados em Cresti 2000; Cresti 2010; Cresti & Moneglia 2010; Panunzi & Scarano 2009) correspondem a núcleos ilocucionários presentes em tipos diferentes de estruturas terminadas (veja 2).

A Figura 1 mostra a frequência de distribuição das diferentes unidades nucleares no minicorpus do C-ORAL-BRASIL. A maior frequência de COM evidencia que na fala espontânea é privilegiada a estratégia comunicativa na qual cada enunciado corresponde a um único ato de fala, ou seja, em que há uma correspondência de um para um entre uma unidade linguística comunicativa

[6] A transcrição completa e o som de todos os exemplos do minicorpus Brasileiro podem ser acessados através da interface de pesquisa do banco de dados IPIC, disponível online.

autônoma e a expressão de uma força ilocucionária, o que denota o caráter pragmático primordial da fala dialógica.

Tabela 1. Unidades que veiculam força ilocucionária em PB

Unidade	**Exemplo**
Comentário COM	Ex. bfamdl04 [103] KAT: copos de quê //=COM=
Comentário Múltiplo CMM	Ex. bfamdl04 [161] SIL: ou é vinho bom caro /=CMM= ou é cerveja //=CMM=
Comentário Ligado COB	Ex. bfammn06 [71] JOR: nós temos vinte-e-cinco funcionários /=COB= dentro de Minas Gerais /=COB= atuando /=COB= com a base nossa aqui em [/1]=SCA= na capital /=COB= e hoje nós tamos /=SCA= numa média de &fature [/1]=SCA= faturamento de um-milhão-e-meio a um-milhão-e-setecentos-mil reais /=SCA= mês //=COM=

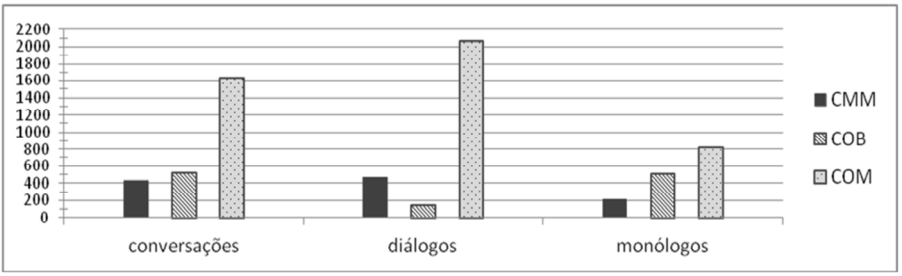

Figura 1. Frequência de distribuição das unidades nucleares de enunciados e estrofes no minicorpus Brasileiro. Fonte: Mittmann & Raso (2012)

Na medida em que há necessidade de uma maior elaboração semântica do discurso falado (monólogo), isso se dá através do aumento na frequência de Comentários Ligados (COB), em estruturas chamadas de *Stanzas*. A fala espontânea monológica é menos ancorada na situação comunicativa imediata e o progresso do discurso é menos afetado pela interação entre os participantes em comparação ao que ocorre na fala espontânea dialógica (Mittmann 2013).

De modo análogo a os enunciados com unidades nucleares de tipo COM, os enunciados cujo núcleo é formado por Comentários Múltiplos (CMM) são mais característicos da fala dialógica (são mais frequentes em diálogos e conversações). Isso ocorre porque a combinação de CMM corresponde a um padrão convencional para expressão de certos valores ilocucionários (Panunzi & Gregori 2011), por natureza, compostos. Estudos com base em corpus da língua italiana evidenciaram ilocuções de reforço, de persuasão, de aproximação ou correção, alternativa ou diretiva dupla, de comparação, de relação necessária, de lista e de clímax (Cresti 2000). Estudo preliminar acerca dos CMM presentes no minicorpus Brasileiro atestou a presença, em PB, de padrões até então não descritos, e também originou uma maior precisão na descrição de padrões já verificados no italiano: ilocução adversativa, pedido de confirmação (desmembrado da diretiva dupla), chamamento funcional, falsa lista, pergunta-resposta e padrões chamados de "duplo-foco" (Arruda 2013).

3.2 As unidades textuais

As unidades de um padrão informacional podem ter funções textuais ou dialógicas. Identificam-se como apresentando funções textuais as unidades informacionais que constroem o texto do enunciado, agregando conteúdo referencial e/ou proposicional. As unidades textuais mais bem descritas até o momento no PB são o Tópico (Mittmann 2012), o Introdutor Locutivo (Maia Rocha 2011; Maia Rocha & Raso 2011) e o Apêndice de Comentário (Oliveira 2012). Estudos preliminares em PB também foram realizados sobre o Parentético e o Apêndice de Tópico (Raso & Ulisses 2008).

A Tabela 2 apresenta uma breve descrição e exemplos das unidades em PB, extraídos do minicorpus Brasileiro. Na sequência, apresentam-se alguns dados sobre as unidades melhor estudadas até o momento em PB. A Tabela 3 mostra a frequência das unidades textuais presentes no minicorpus brasileiro.

O Tópico ocorre com maior frequência em situações comunicativas menos ancoradas no contexto imediato (como os monólogos). Em PB, o TOP é mais comumente realizado através de um sintagma nominal (39,5%) ou verbal (37,7%). No primeiro caso, o SN é tipicamente preenchido por substantivos acompanhados de determinantes, mas há também uma alta ocorrência de pronomes demonstrativos (nos diálogos), e de pronomes pessoais (nos monólogos). No segundo caso, há um uso predominante de SV que expressam situações hipotéticas nas conversações e diálogos, enquanto nos monólogos é mais comum utilizar-se o TOP para delimitar um domínio temporal para o

enunciado. O Tópico é a única unidade informacional, à exceção do Comentário, que apresenta um foco entonacional. O foco delimita o núcleo funcional da unidade e é marcado por um movimento de F0 aliado a um alongamento das sílabas do núcleo. O estudo das características prosódicas do Tópico em PB e em PE revelou a existência de uma realização prosódica do Tópico diferente das observadas no italiano (para detalhamentos da descrição prosódica em do TOP em PB, veja-se Mittmann (2012); para o PE veja-se Rocha (2012); veja-se também Mittmann & Rocha (2012).

Tabela 2. Breve descrição e exemplos das unidades textuais no PB. Baseado em (Cresti 2000; Panunzi & Gregori 2011).

Nome	**Função e exemplo**
Tópico TOP	Fornece uma referência cognitiva para o ato de fala, permitindo que o ato de fala seja distanciado do contexto extralinguístico da situação para que seja ancorado ao contexto estabelecido linguisticamente Ex. bfamdl02 [13] BAL: **as pilhas /=TOP=** eu coloquei aqui //=COM=
Apêndice de Commentário APC	Integração textual do comentário ao qual se refere. Ex. bpubdl02 [86] EUG: tem nenhum par mais /=COM= **daquele modelo //=APC=**
Apêndice de Tópico APT	Informação retardada com relação ao Tópico, adicionar informações mais específicas para o interlocutor. Ex.: bfammn06 [33] JOR: e esse caso /=TOP= **que acontecia /=APT=** marcava muito //=COM=Ex.:
Parentético PAR	Adicionar informação metalinguística, expressar um valor modal. Ex. bfamdl04 [162] SIL: que eu tô ficando /=i-COM= **como se diz /=PAR=** exigente //=COM=
Introdutor Locutivo INT	Assinalar uma mudança de ponto de vista na sequência subsequente, como o discurso reportado. Ex. bfammn04[149] REG: **falei /=INT=** beleza //=COM_r=

Tabela 3. Frequências das unidades informacionais textuais no minicorpus brasileiro

Unidades textuais	**Frequência %**	
Total	**1128**	**100%**
Tópico (TOP, TPL)	600	53%
Introdutor locutivo (INT)	237	21%
Parentéticos (PAR, PRL)	162	14%
Apêndice de Comentário (APC)*	112	10%
Apêndice de Tópico (APT)	17	2%

Chi-quadrado = 883.497 – p < 0,0001 – df =4
Fonte: Mittmann (2012)
*Fonte: Oliveira (2012)

A pesquisa do TOP em PB também revelou a existência de formas particulares de tópicos recursivos, as listas de tópico, que foram depois verificadas também na língua italiana, e o tópico subordinador, não encontrado em italiano (Mittmann 2012). De fato, além das três formas encontradas também no italiano, o PB apresenta uma quarta forma que fora encontrada em PE com frequência aparentemente maior do que em PB.

O Apêndice de comentário posiciona-se após a unidade de COM que integra. A maior parte das expressões em APC corresponde a um conteúdo vazio ou genérico do ponto de vista semântico: repetições de expressões do tema do discurso; preenchimentos, retomada textual e informação tardia. A unidade informacional de APC é geralmente utilizada para realizar a correção ou o acréscimo de material lexical da unidade de COM (Oliveira 2012).

Os Parentéticos constituem-se de inserções que atuam sobre o conteúdo dos eunciados, através das quais passa-se da objetividade à subjetividade da narração ou do ponto de vista (Tucci 2010). Frequentemente, o conteúdo locutivo apresenta elementos modalizadores, que explicitam o juízo pessoal do falante sobre o próprio enunciado, de modo a reforçar ou limitar o conteúdo semântico da ilocução (Tucci 2010; Mello *et al*. 2011). Estudo preliminar do Parentético no PB, realizado com base em quatro textos de aproximadamente 1.500 palavras (dois diálogos e dois monólogos), aponta para a predominância dessa unidade em monólogos e confirma, também para o PB, a prosódia característica da unidade: o PAR é pronunciado com uma frequência fundamental normalmente mais baixa (raramente mais alta) do que o resto do enunciado; a curva melódica é nivelada, mas pode acabar com um movimento ascendente, e há frequentemente aumento na velocidade de elocução do PAR em relação ao resto do enunciado (Vale 2010).

O Introdutor Locutivo sinaliza que o espaço locutivo subsequente apresenta coordenadas espaço-temporais distintas daquelas do nível locutivo primário ancorado à situação enunciativa, instaurando um aqui e agora diferente e sinalizando que tal espaço locutivo não apresenta referências dêiticas válidas para o momento da enunciação. Devido a esse salto hierárquico, as unidades informacionais de Comentário introduzidas pelo INT assumem um valor metailocucionário, que, segundo Cresti, podem ser o discurso reportado, a exemplificação emblemática, o pensamento falado e a narração, podendo ainda também introduzir listas. Através de estudo realizado no PB, a partir de uma amostra de 10 textos do minicorpus Brasileiro, evidenciou-se a existência de outros valores metailocucionários, como a descrição, a instrução, a citação e a contrafactualidade (Maia Rocha 2011; Maia Rocha & Raso 2011). De acordo com o trabalho de Maia Rocha (2011), o INT apresenta como característica acústica um aumento ou abaixamento de sua F0 e aumento da velocidade de elocução da unidade em relação ao restante do enunciado. Em PB, assim como observado no italiano, a curva melódica (movimentos da F0) é descendente e, caso apresente alguma modulação do movimento, o INT tende a terminar com um claro abaixamento da F0, e há sempre um contraste evidente entre a F0 no ponto em que acaba o INT e no ponto em que começa a unidade seguinte. Tal contraste é necessário para que a suspensão pragmática do enunciado seja assinalada.

3.3 As unidades dialógicas

As unidades dialógicas são aquelas que operam sobre a situação comunicativa ou sobre o interlocutor, para garantir o bom funcionamento da comunicação. Alguns estudos já foram realizados sobre algumas unidades dialógicas, em especial sobre os correlatos lexicais e a frequência de ocorrência das unidades em PB e em comparação com outras línguas (Raso & Leite 2010; Raso 2009; Raso in press). Estudos mais aprofundados acerca das propriedades acústicas das unidades dialógicas estão sendo conduzidos no LEEL para melhorar sua descrição e ainda permitir o desenvolvimento de uma metodologia de validação da etiquetagem dessas unidades.

A Tabela 4 apresenta exemplos em PB das unidades dialógicas, extraídos do minicorpus brasileiro, e a Figura 2 mostra a distribuição das unidades dialógicas no minicorpus Brasileiro.

Tabela 4. Exemplos das unidades dialógicas em PB

Nome	Função e exemplo
Incipitário INP	Ex. bfamdl02 [196-197] BEL: ah não / na parte maior / é os + BAL: **não /=INP=** mas é porque eu tô pensando assim //=COM=
Conativo CNT	Ex. bfamcv03 [129] CEL: **olha p' cê ver /=CNT=** vai matar ainda //=COM=
Fático PHA	Ex. bfamcv02 [9] TER: é que ea ganhou tudo /=COM= **né //=PHA=**
Alocutivo ALL	Ex. bfamcv02 [220] RUT: participa não /=COM= **minha filha //=ALL=**
Expressivo EXP	Ex. bfammn05 [35] CAR: **ah /=EXP=** a história dele é muito bonita também //=COM=
Conector Discursivo DCT	Ex.: bpubmn01 [72-73] SHE: então /=PHA= a orientadora /=TOP= ela nũ quer fazer o papel da coordenadora //=COM= **e /=DCT=** vice-versa //=COM=

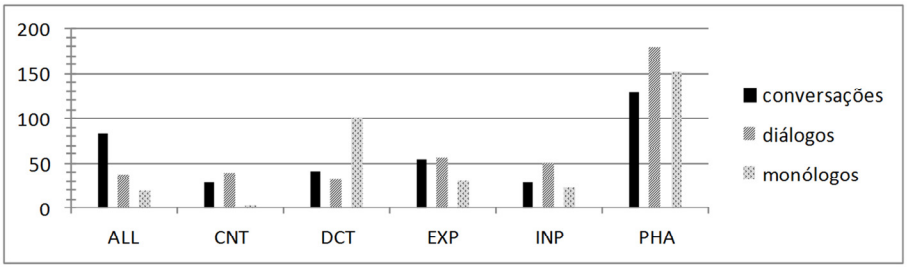

Figura 2. Distribuição das unidades dialógicas no minicorpus Brasileiro. Fonte: Mittmann & Raso 2012

Na sequência, apresenta-se uma breve descrição de cada unidade e suas principais características em PB (Raso in press).

a. Incipitário: é uma marca de contraste afetivo com relação ao enunciado anterior. Pode ter também a função de tomar o turno. Tem contorno

entonacional ascendente-descendente, com alta velocidade de elocução, valor máximo de F0 e de intensidade altos. O preenchimento mais comum em PB é o lexema "não" (40%).

b. Conativo: sua função é induzir o interlocutor a fazer ou parar de fazer algo. Tem contorno entonacional descendente, duração silábica curta e intensidade alta em relação ao resto do enunciado. Ocorre de preferência no início ou final do enunciado. Em PB, os preenchimentos mais comuns são "olha" e variações (50%).
c. Fático: assinala a abertura do canal comunicativo; apresenta um contorno entonacional nivelado ou descendente, com intensidade baixa e duração silábica curta. Sua distribuição no enunciado é livre; em PB, é mais utilizado no fim do enunciado e o preenchimento lexical mais frequente é o lexema "né" (59%).
d. Alocutivo: identifica para quem o ato de fala é dirigido e marca a relação social entre interlocutores. Tem contorno entonacional nivelado ou descendente, com intensidade baixa. Em PB, nomes próprios são os mais frequentes (56%).
e. Expressivo: serve como suporte emocional para o ato de fala, sinaliza coesão social para com o interlocutor. Apresenta muita variação prosódica, provavelmente relacionada ao tipo de suporte emocional e de ilocução. A distribuição é livre, mas preferencialmente ocupa posição inicial. Em PB, as interjeições "ah" (35%) e "nossa" (14%) são os preenchimentos mais comuns.
f. Conector discursivo: une dois enunciados e assinala a continuidade do discurso. O contorno entonacional é nivelado ou modulado, com duração silábica longa e F0 e intensidade altas. Em PB, os preenchimentos mais frequentes são "e" (20%), "porque" (20%), "mas" (16%), "então" (15%) e "aí" (13%).

4. O estudo das Ilocuções e das Atitudes

Uma importante vertente de estudos do projeto C-ORAL-BRASIL é direcionada à identificação e à análise empírica das ilocuções.

L-AcT è uma extensão da teoria dos Atos de Fala de Austin (1962) e, como vimos, define as unidades de referência da fala com base na autonomia pragmática. Isso significa que o núcleo dessas unidades é dado pela ilocução, que pode eventualmente ser acompanhada por unidades não ilocucionárias. Se é

importante o estudo dessas últimas para o conhecimento da estruturação da fala, ainda mais importante é o estudo das unidades nucleares.

A categoria da ilocução, depois de Austin, foi investigada através de paradigmas diferentes. Searle (1969) abriu uma linha de investigação de matriz lógica, que deu muitos frutos, mas também encontrou muitos problemas. Outros autores, como por exemplo, Levinson (1983), se por um lado notam as dificuldades de enfrentar a análise das ilocuções através da dedução lógica, por outros abandonam o âmbito puramente linguístico para uma investigação que insere a ilocução dentro de um quadro comunicativo feito de elementos também extra-linguísticos (Brown & Levinson 1987).

A resposta de L-AcT ao impasse teórico no estudo das ilocuções visa a manter a autonomia da dimensão linguística através da observação empírica. O estudo das ilocuções é visto como estudo das ações verbais que realmente acontecem na comunicação linguística, e que não podem ser deduzidas logicamente, mas somente observadas quando concretamente realizadas e reconhecidas na fala espontânea. Naturalmente, para observar como se realizam as ilocuções na fala é necessário dispor de um corpus, e que ele seja baseado na variação diafásica, ou seja, um corpus no qual seja induzida a realização da maior variedade possível de atos de fala.

Com base na observação de corpora de fala, L-AcT conclui que as ilocuções são veiculadas através da prosódia, a qual constitui a interface entre o ato locucionário e o ato ilocucionário. A prosódia, portanto, se constitui como o veículo necessário das ilocuções. O mesmo material locutivo pode ser realizado prosodicamente de várias maneiras, gerando assim ilocuções diferentes. A sintaxe e a semântica não parecem ser determinantes na realização das ilocuções. Contudo, a prosódia sozinha não pode diferenciar a variedade de ações que se imagina existir na fala humana. Para isso, é necessária a presença de fatores pragmáticos e cognitivos que eliciam a ilocução. Definir esses fatores e estudar a realização prosódica da ilocução são os objetivos de trabalhos que vêm sendo realizados no LEEL. As pesquisas do laboratório LABLITA impostaram a metodologia de estudo das ilocuções da qual partem os estudos do projeto C-ORAL-BRASIL (Cresti 2000, 2012; Firenzuoli 2003; Cresti & Firenzuoli 1999; Moneglia 2011; Rocha 2013).

4.1 Ilocução e atitude

Durante a pesquisa sobre as ilocuções nos deparamos com a grande variação definitória e terminológica sobre essa e outras duas categorias que na literatura

parecem se sobrepor e se confundir entre si e com a ilocução: as categorias de modalidade e de atitude. Uma distinção clara entre ilocução, como categoria pragmática, e modalidade, como categoria semântica, já havia sido formulada com clareza por Cresti (2001). Mello e Raso (2012) retomam as definições de ilocução e modalidade de Cresti (2001), e também a argumentação, com base em testes de comutação, segundo a qual dois elementos que podem ser realizados ao mesmo tempo não devem fazer parte da mesma categoria. De fato, podemos realizar uma determinada ilocução, por exemplo, um convite, com uma modalidade epistêmica ou com uma modalidade deôntica, mas não podemos realizar as duas modalidades ao mesmo tempo, nem um convite contemporaneamente a uma ordem. Através de um experimento, Mello e Raso (2012) tentam mostrar que se a prosódia é essencial em veicular a ilocução, ela, porém, não influi na veiculação da modalidade.

Mello e Raso (2012) observam também a necessidade de reconhecer uma categoria de atitude, definida, parafraseando Bally, como o *Modus* do *Actum*. O argumento é que qualquer ação, ou seja, qualquer ilocução pode ser realizada de maneiras diferentes. Um convite pode ser realizado de maneira sedutora, irritada, apressada, gentil, etc. Naturalmente, não podemos ter mais atitudes ao mesmo tempo, ou seja, não podemos, ao mesmo tempo, realizar a ação em modos diferentes. Para veicular a atitude a prosódia parece essencial tanto quanto o é para a ilocução. Uma pergunta importante é, então, como distinguir os efeitos prosódicos da ilocução e da atitude em um determinado enunciado.

Sempre segundo essa proposta, o domínio sobre o qual a prosódia age para veicular a ilocução não parece ser o mesmo daquele utilizado na veiculação da atitude. A ilocução é veiculada através da marcação prosódica de poucas sílabas da unidade, que constituem o núcleo ilocucionário. As outras sílabas que compõem a locução da unidade se configuram apenas como preparação, coda ou ligação (em caso de núcleos descontínuos). Essas partes servem somente para preencher o conteúdo locutivo, não influindo do ponto de vista prosódico para a veiculação da função ilocucionária. Ao contrário, a atitude parece ter como domínio a unidade inteira. Dependendo da atitude, a marca prosódica parece ser realizada em pontos específicos da unidade ou distribuída na unidade inteira.

Um experimento (Mello & Raso 2012) parece indicar que uma atitude de irritação se distribui na unidade como um todo, pelo menos quando interage com uma ilocução de pedido/convite. Um outro experimento (Rocha 2013) parece, ao contrário, mostrar que uma atitude de cortesia aplicada a uma ilocução de ordem ou de instrução é claramente marcada através de uma subida no final da unidade, independentemente de a sílaba final ser nuclear ou não, ou seja,

independentemente de a sílaba final ser necessária para veicular a ilocução ou não. As manipulações de ordens e instruções coletadas no corpus confirmam essa conclusão. Retirando a subida final de uma ordem ou de uma instrução, obtemos a manutenção da mesma ilocução mas sem a marca de cortesia. O resultado oposto é obtido acrescentando uma subida final a ilocuções de ordem e instruções não marcadas com atitude de cortesia.

Mello e Raso (2012) propõem que a relação recíproca entre as categorias de atitude, ilocução e modalidade não deva ser considerada como uma relação determinística. Em princípio, uma atitude não determina as ilocuções e as modalidades possíveis e *vice-versa*. Contudo, parece haver uma tendência de uma certa atitude "preferir" certas ilocuções e de certas ilocuções "preferirem" certas modalidades, mantendo-se sempre a possibilidade que essa preferência não se realize.

A esse respeito, são interessantes os dados de alguns experimentos realizados no LEEL e ainda não publicados. Os experimentos são relativos a duas ilocuções, oferta e proposta, que em princípio parecem ser veiculadas pelas mesmas formas prosódicas e se distinguem unicamente por fatores pragmáticos e cognitivos. Os experimentos consistiram na eliciação das duas ilocuções em contextos fictícios diferentes, porém com a mesma sequência locutiva. Os contextos de eliciação levam em um caso à interpretação da locução como oferta e em outro como proposta. O experimento foi realizado em três contextos de eliciação para cada ilocução. As curvas prosódicas derivadas mostraram que as duas ilocuções foram realizadas sempre com dois padrões prosódicos distintos. De fato, a forma das duas ilocuções não parece se distinguir, mas a posição dessa forma dentro da unidade é sempre diferente, sendo a de uma ilocução colocada mais à esquerda e a de outra ilocução mais à direita dentro da unidade. Além disso, uma ilocução apresenta parâmetros de F0, duração e intensidade constantemente um pouco superiores à outra. Os testes de substituibilidade mostram que aparentemente a prosódia da oferta é aceitável no contexto de eliciação de proposta, enquanto o inverso parece muito menos aceitável.

Um resultado desse tipo deixa abertos dois caminhos. Devemos considerar as formas prosódicas das duas ilocuções distintas ou devemos pensar que elas normalmente são acompanhadas por duas atitudes distintas que seriam responsáveis pelas diferenças observadas? De fato, parece sensato imaginar que a ilocução de oferta seja feita com um engajamento importante para que a própria ação seja considerada sincera, enquanto a ilocução de proposta não precisa desse engajamento. Isso poderia também explicar o porquê de a prosódia

da oferta ser aceitável no contexto da proposta enquanto o contrário parece muito menos aceitável.

4.2 Ilocuções pesquisadas

Apesar de a reflexão sobre as ilocuções ser fundamental na L-AcT, o grande esforço empreendido para a coleta dos dados que compõem o *corpus* e os estudos sistemáticos sobre as diferentes unidades informacionais não permitiram, por muito tempo, que um estudo sistemático sobre as ilocuções, que se intuía ser muito complexo, pudesse ser feito. O trabalho de Firenzuoli (2003) ainda constitui o máximo esforço de sistematizar uma análise sobre as ilocuções.

Uma retomada dos esforços para sistematizar o estudo sobre as ilocuções é recente, e é devida ao trabalho de Rocha (2013). Até agora no corpus C-ORAL-BRASIL foram estudadas, de forma ainda não definitiva, algumas ilocuções diretivas: a exortação, a proposta, a oferta, a advertência, a ordem e a instrução. A metodologia prevê a realização de alguns passos:

1. coleta do material através do reconhecimento, a partir da escuta sistemática dos textos do corpus, dos enunciados candidatos a representar a ilocução que se deseja estudar;
2. observação mais aprofundada do material coletado para verificar se todos os enunciados coletados realmente veiculam a mesma ilocução;
3. primeira análise das características prosódicas do núcleo; nessa fase, convém analisar exemplos menos afetados por atitudes especialmente marcadas;
4. descrição pragmático-cognitiva do contexto de realização da ilocução. Por características pragmáticas entende-se um restrito número de parâmetros fortemente ligados à situação, como a distância entre os interlocutores ou com relação ao objeto envolvido na ilocução, a possibilidade de os interlocutores se verem ou a visibilidade do objeto envolvido na ilocução, a necessidade ou não de compartilhamento de foco atencional entre os interlocutores, a necessidade de o canal entre os interlocutores estar previamente aberto ou, ao contrário, não estar ainda aberto. Não se trata, portanto, de parâmetros pragmáticos *lato sensu*, mas de elementos situacionais que modificam, no contexto específico, as condições de eliciação. Por fatores cognitivos entende-se a necessidade de os interlocutores compartilharem um determinado conhecimento ou não. Alguns exemplos podem ajudar no entendimento do peso desses parâmetros

(cuja lista é muito reduzida e dos quais um número ainda mais reduzido é pertinentes em um ato específico): a prosódia do chamamento parece sensível ao parâmetro distância e, pelo menos em certas línguas, ao parâmetro visibilidade ou não da pessoa chamada; a prosódia do ato de dêixis (um ato muito frequente e não presente na lista de Searle por não ser reportável a um performativo) parece sensível ao parâmetro movimento do objeto de referência; a prosódia da ordem parece sensível não a uma hierarquia social entre os interlocutores, mas ao controle sobre a situação por parte do falante (possuir um conhecimento que o torna capaz de controlar mais a situação);

5. uma vez que se tem uma primeira descrição prosódica e uma base que justifique uma hipótese sobre os parâmetros pragmáticos e cognitivos, é possível construir uma situação fictícia de eliciação da ilocução e gravar em laboratório a realização da ilocução;
6. a partir de realizações de ilocuções diferentes em contexto fictício mas com o mesmo conteúdo locutivo, é possível fazer a substituição dos áudios e, através de testes de percepção, avaliar a aceitabilidade das substituições e formular hipóteses sobre a compatibilidade ou não da mesma curva para ilocuções diferentes;
7. a última fase consiste na descrição prosódica detalhada e na formulação de hipótese acerca das características que tornam prosodicamente marcadas determinadas ilocuções.

Uma última observação diz respeito aos assim chamados *atos linguísticos indiretos* (Searle 1985). Nossa visão é a de que os atos indiretos, de um ponto de vista estritamente linguístico, não existam. Os atos indiretos convencionalizados (como pedir a hora perguntando se se sabe a hora) são realizados prosodicamente sempre como atos de pedido e não de pergunta. Portanto eles são marcados linguisticamente como pedidos, não necessitando de nenhum processo inferencial para se identificar a força ilocucionária primária. Os atos indiretos não convencionalizados (como, por exemplo, atos de asserção ou reclamação em função do calor, que podem ser interpretados como pedidos para que se abra a porta ou se ligue o ar condicionado) não possuem nenhuma marca linguística que induza a se interpretar o ato como algo diferente do ato realmente proferido. Em contextos específicos esses atos podem desencadear em alguns destinatários processos inferenciais que levam a interpretações comunicativas diferentes. Certamente o processo comunicativo não se limita ao que é marcado linguisticamente, mas é, a nosso ver, fundamental distinguir entre o que é

marcado linguisticamente e o que é interpretado, através de percursos não linguísticos, a partir de algo que, de um ponto de vista linguístico, é marcado de maneira diferente.

5. O estudo da Modalidade

5.1 O que é modalidade?

A modalidade na fala pode ser considerada como a avaliação feita por um falante sobre o material locutivo por ele enunciado – esta visão segue a proposta de Bally (1932), que preconizava a modalidade como a avaliação (*Modus*) sobre algo dito (*Dictum*) por um falante. Entretanto, a definição de modalidade e a classificação de sua tipologia dependem da visão teórica em que um dado estudo se insere. Desta forma, encontram-se desde definições baseadas em visões lógicas (Lyons 1977; Palmer 1986), até aquelas funcionalistas (Bybee & Fleischmann 1995), dentre outras. Nos trabalhos desenvolvidos no escopo do projeto C-ORAL-BRASIL, temos adotado a visão Balliana, implementada por Cresti (2003) e Tucci (2007) em seus estudos, por se tratar de uma abordagem baseada em princípios que observam o uso da fala e suas motivações comunicativas.

5.2 Buscando a Modalidade

A busca por modalidade no C-ORAL-BRASIL enfocou apenas as manifestações lexicais e gramaticais desta categoria semântica e não considerou a modalidade inerente, que requer estudos teóricos ulteriores.

Para uma primeira prospecção de marcação de modalidade, utilizamos o subcorpus informacionalmente anotado. A pesquisa utilizou-se da busca manual por itens modalizadores e da sua classificação em seu contexto de ocorrência de acordo com as seguintes categorias: classe de palavra, unidade informacional de inserção, rótulo semântico (modalidade alética, epistêmica, deôntica), tipologia textual, gênero e nível de escolaridade do falante. Os itens identificados e sua atribuição às categorias supracitadas foram validados através da discussão em grupo entre diferentes anotadores.

Dentre os trabalhos resultantes deste esforço de pesquisa, estão: a identificação de índices morfolexicais de modalidade em unidades informacionais (Mello *et al.* 2010); estudo comparativo entre advérbios de

certeza do português brasileiro e do português europeu (Mello *et al.* 2011), estudo sobre a epistemicidade de construções condicionais[7], estudo sobre as construções condicionais como veículo de modalidade[8], descrição sobre efeitos pragmático-cognitivos de índices modais (Ávila 2012) e o mapeamento de construções adverbiais modais no português brasileiro (Mello & Caetano in press). Adicionalmente aos estudos voltados para a identificação de marcadores de modalidade no C-ORAL-BRASIL, Ávila & Mello (2013) propuseram um esquema de anotação para índices modais na fala espontânea.

Em termos distribucionais, a pesquisa revelou as seguintes observações: em torno de 10% dos 2.573 enunciados examinados, 250, têm algum tipo de marcação modal explícita. A maior parte das marcações (57,85%) é de índices epistêmicos; os índices deônticos correspondem a 23,57% dos casos, enquanto os aléticos perfazem 18,57%. Os índices modais, sua classificação morfolexical e o seu percentual de ocorrência são mostrados na Tabela 5.

A título de ilustração, seguem três exemplos anotados com ocorrências de modalizadores epistêmicos [EPT] e deôntico [DEO]:

(1) =$ [171] não /=PHA= trinta reais /=TOP= aí eu &j [/2]=SCA= eu [/1]=EMP= eu fico imaginando que [EPT] e' fica pensando assim /=INT= Nossa Sio' /=EXP_r= às vezes [EPT] lá em casa tá precisando [DEO] de fazer uma compra e tudo /=COM_r= né //=PHA=$ (bpubmn01)

(2) *LUC: [74] <se [EPT] na primeira vez que cê falou uma palavra /=SCA= não> for /=TOP= nunca mais vai ser [EPT] /=COM= entendeu //=PHA=$ (bfamcv04)

(3) *PAU: [153] porque é capaz [EPT] d' eu subir uma parede lá //=COM=

Na comparação entre advérbios modais de certeza em português brasileiro e europeu, os resultados indicam uma incidência de uso superior em PE que em PB. A hipótese explanatória para tal fato é discutida em Mello *et al.* (2011) e se relaciona às diferenças de hierarquização social e nível de escolarização nas duas culturas. A Tabela 6 apresenta as diferenças numéricas encontradas nas

[7] L. Ávila & P. Côrtes, *A epistemicidade nas construções condicionais do português do Brasil: estudo baseado em corpus de fala espontânea*; trabalho apresentado no XI SILEL, 23-25 novembro 2011, Uberlândia, Brazil.

[8] H. Mello & P. Côrtes, *A transmodular analysis of conditionals in spoken Brazilian Portuguese*, trabalho apresentado no II SILF - Simpósio Internacional de Linguística Funcional, 14-16 agosto 2013, UFSCar, São Carlos, Brasil.

duas variedades da língua portuguesa, distribuídas por tipo interacional e contexto interacional (público e privado).

Tabela 5. Índices modais, tipologia e percentuais de ocorrência

Classes morfolexicais	Tipos	Percentuais
Adjetivos (ou nominais em função adjetival) em posição predicativa	(é) lógico, é provável, é importante, (é) verdade	1,42%
Advérbios e expressões adverbiais	talvez, certamente, realmente, às vezes, também, logicamente, sinceramente, com certeza, completamente, sem dúvida, possivelmente, na verdade, na realidade	6,42%
Condicionals	[se X então Y]	13,21%
Construções verbais modalizadoras	tem condição (de), tem chance de, o que acontece, ter que, ficar imaginando, ficar pensando, (é) para + inf., dá para + inf., ter certeza, vai saber, tem jeito	22,14%
Futuro composto	vou + inf.	1,07%
Futuro do Pretérito	ia ser, ia dar, seria	3,21%
Outras formas	digamos que, de certa forma	3,57%
Verbos (modo indicativo – presente, perfeito e imperfeito; infinitivo)	dever, poder, achar, acreditar, acontecer, ver, conseguir, precisar, pensar, dar e parecer.	48,92%

Tabela 6. Advérbios modais de certeza em PE e PB

	Público EP/BP	Privado EP/BP	TOTAL EP/BP
Monólogos	26/5 (5.2)	23/8 (2.875)	49/13 (3.77)
Diálogos	36/25 (1.44)	11/8 (1.375)	47/33 (1.424)
Conversações	23/6 (3.83)	22/8 (2.75)	45/14 (3.214)
TOTAL	85/36 (2.36)	46/24 (1.916)	141/60 (2.35)

O estudo sobre advérbios modalizadores, feito no corpus C-ORAL-BRASIL completo (Mello & Caetano in press), mostra os seguintes resultados: um total de 763 exemplares, divididos em 28 tipos distintos, com alta concentração em um único tipo – o advérbio *mesmo* (55%). A busca por advérbios foi realizada através da anotação PoS executada pelo *parser* Palavras (Bick 2000, 2012), manualmente conferida para precisão e acurácia. Exceto por um único tipo deôntico - *necessariamente* – todos os demais são advérbios epistêmicos. A alta frequência de *mesmo* demanda um estudo à parte para a certificação de sua função e situação de uso.

O estudo de construções condicionais e sua epistemicidade foi conduzido a partir dos dados do minicorpus do C-ORAL-BRASIL (Mello & Côrtes 2013). Foi encontrada a seguinte distribuição, na Tabela 7, das construções condicionais nos dados analisados:

Tabela 7. Frequência de construções condicionais

Tipologia de interação	Contexto	Frequência
Monólogo	Privado	18
	Público	6
Diálogo	Privado	27
	Público	13
Conversação	Privado	38
	Público	9

A distribuição das construções condicionais em relação à sua estrutura sintática pode ser vista na Tabela 8.

Tabela 8. Distribuição das estruturas sintáticas de construções condicionais

Estrutura Sintática	Frequência
Protáse- Apódose	75
Apódose- Protáse	12
Protáse	24

Neste estudo, as condicionais foram encontradas distribuídas em variadas estruturas informacionais: prótase e apódose na mesma unidade de comentário, no perfil tópico/comentário, em enunciados distintos, em perfis de comentários múltiplos etc. Quando a prótase e a apódose são realizadas em enunciados

diferentes, a estrutura se completa em atos de fala diferentes, cada um veiculando uma ilocução.

6. Estudos em andamento

Entre os outros estudos em andamento há quatro projetos em estado mais avançado que merecem ser expostos de maneira mais detalhada.

6.1 Compilação do corpus formal de PB

O primeiro dos estudos em andamento é a realização do corpus C-ORAL-BRASIL II, que, uma vez completado, consistirá de textos formais em contexto natural, textos telefônicos e textos de mídia televisiva. Quando esse segundo corpus estiver completo, a comparabilidade entre o C-ORAL-ROM e o C-ORAL-BRASIL será integral. Como para o C-ORAL-BRASIL I (informal), o objetivo é o de compilar um corpus que seja cerca de 30% maior do que os corpora do C-ORAL-ROM. Há uma única diferença significativa em relação aos corpora C-ORAL-ROM: decidiu-se pela não inserção das gravações de interações entre homem e máquina e pela ampliação da quantidade de gravações telefônicas. Portanto, a arquitetura desta segunda metade do corpus prevê, em número de palavras, as seguintes proporções, de maneira a planejar a comparabilidade principalmente com o corpus italiano: cerca de 42% de contexto natural formal, cerca de 17% de textos telefônicos e cerca de 41% de mídia, para um total de aproximadamente 180.000 palavras. O contexto natural será dividido, em partes aproximadamente iguais, entre os seguintes contextos: discursos políticos, debates políticos, religião, conferências, explicações profissionais, negócios, direito e ensino. Contrariamente ao corpus informal, mas de acordo com os textos formais do C-ORAL-ROM, a medida de cada texto será em princípio maior, com uma média de cerca de 3.000 palavras cada um. O contexto mídia, de acordo com o C-ORAL-ROM, compõe-se de textos pertencentes aos seguintes gêneros: entrevistas (cerca de 15%), previsões do tempo (0,5%), notícias (3%), esporte (10,5%), imprensa científica (11,5%), reportagens (20%) e *talk-shows* (40%). Nesse caso, a dimensão dos textos muda dependendo do gênero, variando de textos muito pequenos, como no caso das previsões de tempo, a textos de mais de 3.000 palavras em outros gêneros. Também os textos telefônicos variam de tamanho, sem nunca ultrapassar as 3.000 palavras. No momento, quase todas as gravações do contexto natural e de telefone já foram realizadas, assim como a transcrição da maioria dos textos

gravados e a revisão de parte dos textos. A previsão é que o corpus C-ORAL-BRASIL II seja disponibilizado em 2016.

6.2 Etiquetagem informacional de um minicorpus de inglês americano

O segundo projeto em andamento prevê a etiquetagem informacional de um minicorpus de inglês americano. Para essa finalidade foram escolhidos 22 textos do *Santa Barbara Corpus* (Du Bois *et al.* 2000). Os textos foram escolhidos de maneira a permitir a comparabilidade com os dois minicorpora etiquetados para o italiano e o PB. Esses textos já foram resegmentados e alinhados com base nos critérios da família C-ORAL. A primeira etiquetagem está sendo completada; em momento posterior será necessária uma fina revisão da etiquetagem antes de serem iniciados os estudos. Contudo, exemplos extraídos do minicorpus Santa Barbara já foram utilizados para uma exposição em inglês da L-AcT (Moneglia & Raso in press).

Vale a pena observar que durante a etiquetagem do minicorpus inglês foi observada uma frequência muito alta de formas de tópico ausentes na observação do italiano e, observadas pela primeira vez, em PB e posteriormente, com frequência ainda maior, em PE. A forte frequência dessa forma em PE e em inglês, comparada com uma presença claramente menos frequente em PB e com a aparente ausência em italiano nos levou a supor uma correlação, que deve ser verificada, entre o tipo de forma prosódica para o tópico e o estatuto mais ou menos acentual de uma língua.

6.3 As formas pronominais você(s)/ocê(s)/cê(s) em PB

O terceiro projeto em andamento, já bastante avançado, é uma pesquisa promissora sobre os pronomes *você/ocê/cê*. Trata-se de um assunto já muito pesquisado, mas com premissas teóricas e metodologias muito diferentes. Em linha geral, as pesquisas sustentam que a forma *você* e *ocê* são sempre tônicas e a forma *cê* sempre clítica (ou fraca). Essa conclusão é alcançada deduzindo a tonicidade ou a cliticidade a partir das funções sintáticas (supostamente) gramaticais e não-gramaticais, ou através de análises acústicas feitas a partir da realização de frases em laboratório. O estudo em curso, e já parcialmente publicado (Ferrari 2013), usa uma metodologia nova, possível unicamente a partir de um corpus com as características de coleta de dados, de transcrição e de alinhamento do C-ORAL-BRASIL. Em primeiro lugar, a pesquisa se baseia na óbvia constatação de que tonicidade e atonicidade são propriedades acústicas. Elas devem, portanto, ser definidas e demonstradas através de um exame

fonético. Ao mesmo tempo, não é pensável que o uso real seja demonstrável através do exame acústico de frases realizadas em um contexto controlado, sem que se realizem as condições próprias da comunicação natural.

No corpus foram encontradas as diferentes formas de acordo com a seguinte frequência: 2137 *cê*, 125 *cês*; 281 *ocê*, 31 *ocês*; 312 *você*, 39 *vocês*. Portanto, 77% das ocorrências do corpus são relativas à forma *cê(s)*. Essa porcentagem permanece invariada mesmo retirando-se as repetições em casos de *retractings*. Dos cê(s) que sobram depois da retirada das repetições, 2075 ocupam uma posição pré-verbal com função de sujeito, 36 ocupam uma posição pós-verbal e 31 ocupam uma posição pré-verbal, porém isolada em unidade tonal. Todos os casos de posição pós-verbal são preposicionados, com as formas preposicionais *com*, *c'*, *p'* e *d'*. Esses casos representam cerca de um quarto de todos os objetos preposicionados do corpus nas formas pronominais de segunda pessoa. É difícil sustentar que tal porcentagem seja fruto de problemas de execução. O interessante é que medindo-se a duração dessas formas, observa-se que elas têm valores comparáveis com aqueles das sílabas mais longas dos enunciados aos quais pertencem, enquanto as formas de *cê(s)* pré-verbais apresentam valores comparáveis com aqueles das sílabas mais curtas (vejam-se as análises acústicas em Ferrari 2013). Entre as formas isoladas em unidade tonal, uma parece ter a função de tópico e de fato é extremamente longa. Contudo, em se tratando de um único exemplo, não se pode afirmar com certeza que a forma *cê* pode realizar essa função. Nesse estágio do trabalho, podemos contudo afirmar que:

1. a forma *cê* pode (contrariamente ao que é afirmado na literatura) realizar um objeto preposicionado;
2. a forma *cê* pode ser tanto clítica quanto tônica; ou seja, a tonicidade do pronome não pode ser deduzida a partir da sua forma segmental.

Essas duas afirmações são exemplificadas nos dois exemplos a seguir (Figura 3 e Figura 4).

(1) bfamcv14 [153] apresenta um cê pré-verbal muito curto

A duração das sílabas do enunciado, depois da normalização, apresentam os seguintes valores: se -3.32; pa -0.04; sA -1.02; ma -0.63; zonS -1.04; peh -0.63; gA -0.89; koN -0.84; toRn -1.53; segI 3.75; pa -0.89; sa -2.68; ma 0.39; zonS 4.60. A primeira sílaba, como se vê, apresenta, depois da normalização (ou seja,

depois que foi retirada a influência da duração intrínseca do nível segmental) o valor claramente mais curto entre as sílabas do enunciado[9].

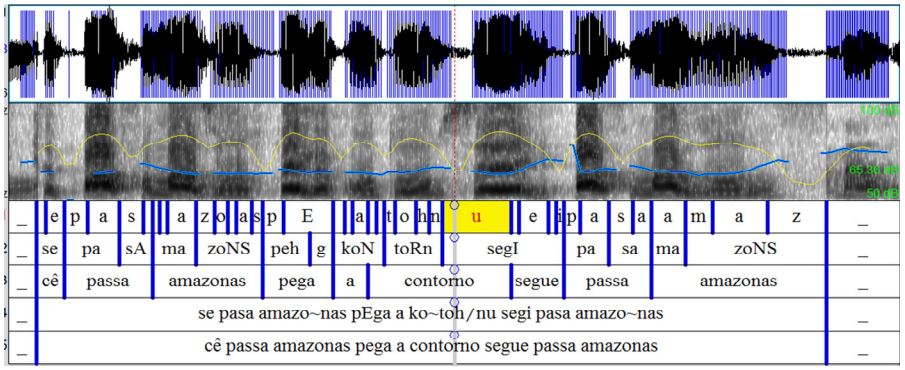

Figura 3. Análise do exemplo 1

(2) bpubcv02 [266] apresenta um cê pós-verbal, não final, muito longo

A duração das sílabas do enunciado, depois da normalização, apresentam os seguintes valores: IN 1.98; taU -1.97; fi -0.94; ka -2.23; koN -3.47; se -0.78; oS -2.10; vaU -2.26. Como pode-se notar, apesar de os valores serem quase todos negativos, apenas uma sílaba (a inicial) é mais longa do que o *cê* pós-verbal, que definitivamente não pode ser considerado curto, e portanto átono.

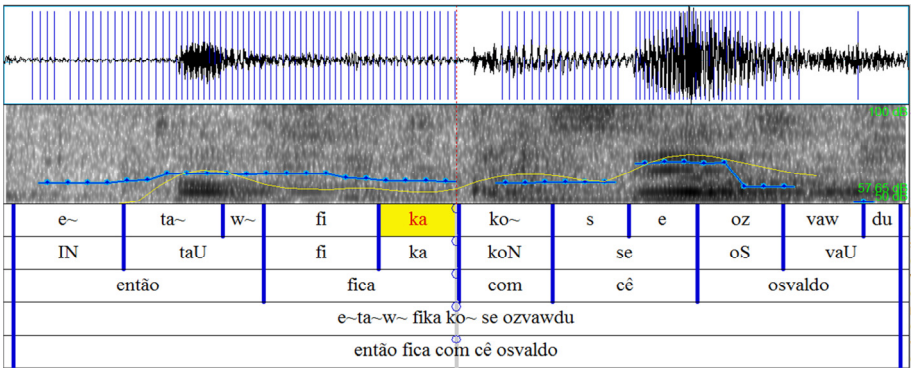

Figura 4. Análise do exemplo 2

[9] A normalização foi obtida utilizando o script SGDetector (Barbosa 2013).

6.4 Validação das unidades dialógicas

Em 3 foi antecipado que está em curso uma análise prosódica das unidades dialógicas. O objetivo dessa análise é validá-las estatisticamente. A importância desse trabalho é devida também ao fato de as unidades dialógicas (UD) do arcabouço da L-AcT constituírem uma explicação relativa aos Marcadores Discursivos (MD). De fato, nenhuma teoria sobre os MDs consegue responder às duas perguntas seguintes: 1. como prever quando um item lexical (ou um pequeno conjunto de itens lexicais) é um MD e quando não é um MD? 2. Uma vez estabelecido que um item é um MD, como prever a sua função específica?

À primeira pergunta é possível responder identificando na quebra prosódica a interrupção da composicionalidade sintática. Desta forma, existe uma marca previsível que diferencia um item lexical composicional com o resto do enunciado de um item não composicional. O item não composicional, contudo, nem sempre se constitui como MD. De fato, se a quebra prosódica for de natureza terminal, o item funciona como uma ilocução (isso impede, portanto, a afirmação de que as interjeições se constituem sempre como DM).

Um exemplo pode esclarecer a situação (para maiores detalhes, veja-se Raso (in press).

(3) *PAU: ah / não acaba não / acaba // (bpubdl01-119)

(4) *PAU: ah / não / ea disse que é pa ficar / por algum tempo // (bpubdl01-197)

(5) *ROG: aqui já tá dando [/4] aqui já tá dando a altura //

(6) *PAU: olha aqui + não // tá dando a altura / daquele que a <Isa> marcou <lá> / né // (bpubdl01-12-14)

Nos exemplos temos três ocorrências de *não* marcadas em itálico. No primeiro o não é composicional com o verbo e, portanto, não é um candidato a ser considerado como MD. O segundo *não* é precedido e seguido por uma quebra prosódica que impede a composicionalidade, e é, portanto, um bom candidato. O terceiro, também está isolado entre quebras prosódicas e poderia, assim, ser um bom candidato. Contudo, no terceiro caso o *não* é interpretável pragmaticamente em isolamento, enquanto no segundo a interpretação não é possível. Isso pode ser verificado escutando-se os áudios dos exemplos. O *não* interpretável será classificado como ilocucionário, enquanto o não-interpretável como UD ou MD.

Vimos assim como é possível responder à primeira pergunta com base em critérios prosódicos. Responder à segunda pergunta é mais difícil. Como antecipado de forma simplificada em 3, a teoria, a partir da observação dos corpora, identificou 6 tipos de UD (ou MD) com características prosódicas diferentes e que veiculam cada um uma função diferente. O trabalho em curso está medindo as características prosódicas de todas as UDs etiquedas no minicorpus com o objetivo de verificar se é possível demonstrar estatisticamente a existência de 6 agrupamentos.

Existem muitas dificuldades metodológicas a serem enfrentadas. A mais importante é que na fala espontânea não temos um termo de comparação fixo para avaliar as medidas. Portanto cada unidade deve ser avaliada em relação ao enunciado em que aparece. Mesmo assim, a estruturação dos enunciados é extremamente variável, não permitindo que um enunciado seja comparável com outro. O que se resolveu foi tomar como termo de comparação para as medidas sempre a ilocução do enunciado ou a mais próxima de uma estrofe. Isso apenas reduz a variabilidade, uma vez que os comentários também são variáveis. Para limitar ainda mais os efeitos da variabilidade, a observação de situações idiossincráticas (um comentário especialmente alongado ou excepcionalmente intenso ou com um movimento de F0 extremo) permite isolar casos que fogem de uma média confiável.

Como os parâmetros acústicos que entram em jogo são muitos (não somente os principais indicados em 3), o trabalho de mensuração e, depois, de validação estatística é muito complexo.

6.5 Outros estudos

Entre os outros estudos em andamento mencionamos apenas um estudo que discute a relação entre pausas e quebras prosódicas, um estudo sobre a negação no PB e dois estudos sobre a subordinação na fala, os últimos três ainda em estádio inicial.

Está sendo completado um estudo que discute a relação entre pausas e quebras prosódicas para a segmentação da fala. Foram escolhidas randomicamente 10 sequências de enunciados do mesmo falante em cada texto do corpus, com o objetivo de medir a existência ou não de pausas (e suas durações) tanto em fronteiras de enunciados quanto dentro de enunciados. Os resultados serão submetidos a testes estatísticos. O objetivo do trabalho é mostrar que, qualquer que seja a duração temporal estabelecida para a pausa, haverá sempre certa quantidade de fronteiras entre enunciados não identificáveis

através da pausa e, ao contrário, haverá sempre certa quantidade de enunciados onde será identificada, apenas com base na pausa, uma ou mais fronteira internas e que, portanto, serão divididos em dois ou mais enunciados. Naturalmente, quanto maior for a duração estabelecida para a medida de pausa, tanto menor será o risco de dividir internamente os enunciados, mas tanto maior será aquele de não identificar as fronteira entre enunciados diferentes, e vice-versa. Possuir medidas estatísticas precisas irá ajudar em uma discussão ainda permanece aberta sobre a segmentação da fala.

Outro estudo em andamento enfoca as restrições para o uso das três estruturas de negação no PB: a negação pré-verbal, a negação pós-verbal e a negação dupla. Nesse estudo se prestará atenção também a eventuais efeitos devidos a fenômenos de redução fonética da negação. Esse estudo é de importância considerável para a descrição de um traço do PB conhecido e também estudado, porém através de metodologias introspectivas ou em corpora que não podem ser considerados espontâneos.

Finalmente, estão sendo estudadas as estruturas de subordinação da fala, principalmente as relativas e as completivas. O modelo destes estudos sobre o PB é constituído por trabalhos recentes de Emanuela Cresti, principalmente Cresti (in press). O grande interesse desses trabalhos não está apenas em quantificar os dois principais tipos de subordinação da fala, mas em observar como a subordinação interage com a segmentação prosódica e a etiquetagem informacional. De fato, a sintaxe da fala espontânea é definida pela L-AcT segundo sua distribuição no contínuo da fala, ou seja, nas 'ilhas' que o compõem e, portanto, assume a importância de distinguir duas tipologias de sintaxe:

a. linearizada (*linearized syntax*): estruturas de coordenação e subordinação próprias dentro da mesma unidade tonal, ou seja, dentro da mesma 'ilha'; um exemplo dessa tipologia é: *GIL: eu nũ sei que cês acham //=COM=
b. articulada (*patterned syntax*): estruturas sintáticas realizadas em mais unidades tonais, ou seja, ao longo de várias 'ilhas', cada uma das quais desenvolve uma função informacional diferente[10]; um exemplo dessa tipologia é: *LUZ: porque eu acho que no mesmo concurso/=TOP= cê nũ pode fazer duas //=COM=

[10] Cresti (in press); Miller e Weinert (1998: 22) falam de *integrated* vs. *fragmented* ou *unintegrated syntax*. Nos termos de Blanche-Benveniste (2000), a distinção entre sintaxe linearizada e articulada corresponde à distinção entre *micro-* e *macro-syntaxe*.

Entender os efeitos que os dois tipos de subordinação realizam na fala é o objetivo desses estudos.

Agradecimento

Tommaso Raso e Heliana Mello agradecem o empenho de toda a equipe do C-ORAL-BRASIL no seu desenvolvimento. Apoio ao projeto C-ORAL-BRASIL tem sido mantido pelo CNPq, FAPEMIG e UFMG.

Referências

Arruda, A. 2013. *A unidade informacional de Comentário Múltiplo em Português Brasileiro: uma análise baseada em corpus*. BA Thesis, Universidade Federal de Minas Gerais.

Austin, J.L. 1962. *How to Do Things with Words*. Oxford: Oxford Univerity Press.

Ávila, L. 2012. Pensar a modalidade: marcação epistêmica e evidencial no português brasileiro falado. In D. Dutra & H. Mello (eds), *Atas do X Encontro de Linguística de Corpus*. Belo Horizonte: FALE-UFMG, 37-56.

Ávila, L. & Mello, H. 2013. Proposta para um esquema de anotação da modalidade em um minicorpus oral do Português Brasileiro: desafios preliminares. In É. Laporte, A. Smarsaro & O.A. Vale (eds), *Dialogar é preciso: linguística para o processamento de línguas*. Vitória: PPGEL/UFES, vol. 1, 31-44.

Bally, C. 1932. *Linguistique générale et linguistique française*. Berna: Francke Verlag.

Barbosa, P. A. 2013. Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. In B. Bigi & D. Hirst (eds), *Proceedings from TRASP 2013*, Laboratoire Parole et Langage, Aix-en-Provence, 86-90.

Berruto, G. 1987. *Sociolinguistica dell'Italiano Contemporaneo*. Roma: La Nuova Italia Scientifica.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. & Conrad, S. 2001. Register variation: A corpus approach. In D. Schiffrin, D. Tannen & H. Hamilton (eds), *The Handbook of Discourse Analysis*. Oxford: Blackwell, 175-196.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press

Bick, E. 2000. *The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Bick, E. 2012. A anotação gramatical do C-ORAL-BRASIL. In T. Raso & H. Mello, *C-ORAL-BRASIL I Corpus de referência do português brasileiro falado informal* (eds). Belo Horizonte: UFMG, 223-254.

Blanche-Benveniste, C., Bilger, M., Rouget, C., Van den Eynde, K. & Mertens, P. 1990. *Le français parlé*: *Études grammaticales*. Paris: C.N.R.S.

Blanche-Benveniste, C. 2000. Le français parlé: un regard sur sa syntaxe. In G. Antoine G. & B. Cerquiglini (eds), *Histoire de la langue française 1945-2000*. Paris: CNRS Editions, 195-197.

Brown, P. & Levinson, S. 1987. *Politeness*: *Some universals in language usage*. Cambridge: Cambridge University Press.

Bybee, J. & Fleischmann, S. (eds) 1995. *Modality and grammar in discourse*. Amsterdam/Philadelphia: John Benjamins.

C-ORAL-BRASIL. http://www.c-oral-brasil.org (acessed March 15, 2014).

C-ORAL-ROM. http://lablita.dit.unifi.it/coralrom (acessed March 15, 2014).

CorpusEye. http://corp.hum.sdu.dk/cqp.pt.html (accessed March 15, 2014).

Cresti, E. 2000. *Corpus di Italiano parlato*. Firenze: Accademia della Crusca.

Cresti, E. 2001. Modalità e illocuzione. In P. Beccarla & C. Marello (eds), *Scritti in onore di Bice Mortara Garelli*. Torino: Edizioni dell'Orso.

Cresti, E. 2003. Ilocution et modalité dans le comment et le topic. In A. Scarano (ed.), *Macrosyntaxe et pragmatique: la analyse linguistique de l'oral*. Roma: Bulzoni, 133-182.

Cresti, E. 2010. La Stanza: un'unità di costruzione testuale del parlato. In A. Ferrari (ed.), *Atti del X Congresso SILFI. Sintassi storica e sincronica nell'italiano*. Firenze: Franco Cesati Editore, 713-732.

Cresti, E. 2012. The definition of Focus in Language into Act Theory. In H. Mello, A. Panunzi & T. Raso (eds), *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*. Firenze: FUP, 39-82. http://www.fupress.com/archivio/pdf/5030.pdf

Cresti, E. in press. Syntactic properties of spontaneous speech in the Language into Act Theory: data on Italian complements and relative clauses. In T. Raso & H. Mello (eds), *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins.

Cresti, E. & Firenzuoli, V. 1999. Illocution et profils intonatifs de l'italien. *Revue Française de Linguistique Appliquée* IV(2): 77-98.

Cresti, E. & Gramigni, P. 2004. Per una linguistica corpus based dell'italiano parlato: Le unità di riferimento. In F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy (eds), *Atti del Convegno Nazionale "Il Parlato Italiano"*, CD-ROM. Napoli: M. D'Auria, 1-26.

Cresti, E. & Moneglia, M. 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages.* Amsterdam/Philadelphia: John Benjamins.

Cresti, E. & Moneglia, M. 2010. Informational patterning theory and the corpus-based description of spoken language: The compositionality issue in the topic-comment pattern. In M. Moneglia & A. Panunzi (eds), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective.* Firenze: Firenze University Press, 13-45. http://digital.casalini.it/9788884535290.

DB-IPIC. http://lablita.dit.unifi.it/app/dbipic (accessed March 15, 2014).

DeMauro, T., Mancini, F., Vedovelli, M. & Voghera, M. 1993. *Lessico di Frequenza dell'Italiano Parlato.* Milano: ETAS.

Du Bois, J.W., Chafe, W.L., Meyer, C. & Thompson, S.A. (orgs) 2000. *Santa Barbara Corpus of Spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Ferrari, L. 2013. As formas cê(s) e você(s) na fala espontânea do PB: uma análise baseada em corpora. Domínios de Linguagem (7:2), 200-237. http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/23735

Firenzuoli, V. 2003. Forme Intonative di Valore Illocutivo dell'Italiano Parlato: Analisi Sperimentale di un Corpus di Parlato Spontaneo. PhD diss., Università degli Studi di Firenze.

Gadet, F. 1996a. Niveaux de langue et variation intrinsèque. *Palympsestes* 10, 17-40.

Gadet, F. 1996b. Variabilité, variation, variété. *Journal of French Language Studies* 1, 75-98.
Gadet, F. (ed.) 1997. *La Variation en Syntaxe. Langue Française* 115 (special issue).
Gadet, F. 2000. Vers une sociolinguistique des locuteurs. *Sociolinguistica* 14, 99-103.
Gadet, F. 2003. *La Variation Sociale en Français*. Paris: Ophrys.
Givón, T. (org) 1979. *Discourse and Syntax*. New York: Academic Press.
Gregori, L. & Panunzi, A. 2012. DB-IPIC: an XML database for informational patterning analysis. In H. Mello, M. Pettorino & T. Raso, *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*. Firenze: Firenze University Press, 121-125. http://www.fupress.com/catalogo/proceedings-of-the-viith-gscp-international-conference/2417.
Halliday, M.A.K. 1989. *Spoken and Written Language.* Oxford: Oxford University Press.
Levinson, S. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
Linguateca. http://www.linguateca.pt/ACDC/ (accessed March 15, 2014).
Lyons, J. 1977. *Semantics*. Cambridge: Cambridge University Press.
MacWhinney, B. 2000. *The CHILDES Project*: Tools for Analyzing Talk. Vol 1: The Format and Programs. Mahwah, NJ: Lawrence Erlbaum Associates. Third Edition.
Maia Rocha, B. 2011. *A Unidade Informacional de Introdutor Locutivo no Português Brasileiro: uma análise baseada em corpus.* MA Thesis, Universidade Federal de Minas Gerais, Belo Horizonte. http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/DAJR-8ELJXZ
Maia Rocha, M. & Raso, T. 2011. Estudo contrastivo do uso de alocutivos em português brasileiro, italiano e em falas de italianos bilíngues em contato prolongado com o português do Brasil. *Revista de Italianística* (21-22), 53-64. http://www.revistas.usp.br/italianistica/issue/view/5473.
Mello, H., Carvalho, J. & Côrtes, P. 2010. Modalidade na fala espontânea do português brasileiro: um primeiro mapeamento de índices morfolexicais. *Revista Estudos da Linguagem* v. 18, n. 2, 105-133.
Mello, H., Panunzi, A. & Raso, T. (eds) 2012. *Pragmatics and Prosody. Illocution, Modality, Attitude, Information Patterning and Speech Annotation*. Firenze: Firenze University Press.
Mello, H., Pettorino, M. & Raso, T. (orgs) 2012. *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*. Firenze: Firenze University Press. http://www.fupress.com/catalogo/proceedings-of-the-viith-gscp-international-conference/2417
Mello, H.R., Ramos, A.C. & Avila, L.B.B. 2011. Probing modal adverbs in Brazilian and European Portuguese: sociocultural variability in a pluricentric language. In A.S. Silva, A. Torres & M. Gonçalves (eds), *Pluricentric Languages: Linguistic Variation and Sociocognitive Dimensions*. Braga: Aletheia, 473-486.
Mello, H. & Raso, T. 2012. Illocution, Modality, Attitude: Different names for different categories. In H. Mello, A. Panunzi & T. Raso (eds), *Pragmatics and Prosody. Illocution, modality, attitude, information patterning and speech annotation*. Firenze: Firenze University Press, 1-18. http://www.fupress.com/archivio/pdf/5030.pdf
Mello, H. & Caetano, R. in press. Mapeamento de advérbios modalizadores no português brasileiro falado.
Mello, H. & Raso, T. 2009. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades* v. 13, n. 1, 153-178.
Mello, H., Raso, T., Mittmann, M., Vale, H. & Côrtes, P. 2012. Transcrição e segmentação prosódica do *corpus* c-oral-brasil: critérios de implementação e validação. In T. Raso &

H. Mello (eds), *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, 125-176.

Miller, J. & Weinert, R. 1998. *Spontaneous Spoken Language*. Oxford: Clarendon Press.

Mittmann, M. 2012. O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o Tópico no Português Brasileiro. PhD diss., Universidade Federal de Minas Gerais, UFMG, Belo Horizonte. http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/LETR-97YMKT

Mittmann, M. 2013. Análise da estruturação de diálogos e monólogos na fala informal: quantificando as diferenças. *Domínios de Lingu@gem* 7(2), 338-372. http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/23760/13584

Mittmann, M. & Raso, T. 2012. The C-ORAL-BRASIL informationally tagged mini-corpus. In H. Mello, A. Panunzi & T. Raso (eds), *Pragmatics and Prosody. Illocution, modality, attitude, information patterning and speech annotation*. Firenze: Firenze University Press, 151-183. http://www.fupress.com/archivio/pdf/5030.pdf

Mittmann, M. & Rocha, B. 2012. Prosodic features of Topic Information Unit in BP and EP: a corpus based study. In H. Mello, M. Pettorino & T. Raso (eds), *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*. Firenze: Firenze University Press, 202-206. http://www.fupress.com/archivio/pdf/2417_5900.pdf (accessed on May 15, 2014)

Moneglia, M. & Martin, Ph. 2005. The C-ORAL-ROM resource. In E. Cresti & M. Moneglia (eds), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins, 1-70.

Moneglia, M. 2011. Spoken Corpora and Pragmatics. *Revista Brasileira de Linguística Aplicada* (2), 479-519. http://www.scielo.br/pdf/rbla/v11n2/a09v11n2.pdf

Moneglia, M. & Cresti, E. 1997. Intonazione e criteri di trascrizione del parlato. In U. Bortolini & E. Pizzuto (eds), *Il Progetto CHILDES Italia: Contributi di Ricerca sulla Lingua Italiana*. Pisa: Del Cerro, 57-90.

Moneglia, M. & Raso, T. in press. Notes on Language into Act Theory. In T. Raso & H. Mello (eds), *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins.

Nencioni, G. 1983. *Di scritto e di parlato: discorsi linguistici*. Bologna: Zanichelli.

Oliveira, C. 2012. O Apêndice de Comentário no Português do Brasil - uma análise baseada em corpus. PhD diss., Universidade Federal de Minas Gerais. http://www.letras.ufmg.br/poslin/defesas/1288D.pdf

Palmer, F.R. 1986. *Mood and Modality*. Cambridge: Cambridge University Press.

Panunzi, A. & Gregori, L. 2011. DB-IPIC: An XML Database for the representation of information structure in spoken language. In H. Mello, A. Panunzi & T. Raso (eds), 133-150. http://www.fupress.com/Archivio/pdf\5030.pdf.

Panunzi, A. & Mittmann, M. in press. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In T. Raso & H. Mello (eds), *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins.

Panunzi, A. & Scarano, A. 2009. Parlato spontaneo e testo: analisi del racconto di vita. In L. Amenta & G. Paternostro (eds), *I parlanti e le loro storie. Competenze linguistiche, strategie comunicative, livelli di analisi: Atti del Convegno Carini-Valderice*. Palermo: Centro di studi filologici e linguistici siciliani, 121-132.

Praat: doing phonetics by computer. http://www.praat.org/ (accessed March 15, 2014).

Raso, T. 2009. Estudo contrastivo do uso de alocutivos em português brasileiro e italiano. *Fragmentos* (37), 145-163.

https://periodicos.ufsc.br/index.php/fragmentos/article/download/27393/24626
Raso, T. 2012. O *corpus* C-ORAL-BRASIL. In T. Raso & H. Mello (eds), *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: UFMG, 55-90.
Raso, T. 2013. Fala e escrita: meio, canal, consequências pragmáticas e linguísticas. *Domínios de Linguagem* (7:2), 12-46. http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/23730/13568
Raso, T. in press. Prosodic constraints for Discourse Markers. In T. Raso & H. Mello (eds), *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins.
Raso, T. & Leite, F. 2010. Estudo contrastivo do uso de Alocutivos em italiano, português e espanhol europeus e português brasileiro. *Domínios de Lingu@gem*, (1). http://www.dominiosdelinguagem.org.br/pdf/dl7/DL7-10.pdf
Raso, T. & Mello, H. (eds) 2012. *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: UFMG.
Raso T. & Mello, H. (eds) in press. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins.
Raso, T. & Mittmann, M. 2009. Validação estatística dos critérios de segmentação da fala espontânea no *corpus* C-ORAL-BRASIL. *Revista de Estudos da Linguagem* 17:2, 73-91. http://relin.letras.ufmg.br/revista/upload/17-2_04.pdf
Raso, T. & Ulisses, A. 2008. Tópico e apêndice no português do Brasil: algumas considerações. *Revista de Estudos da Linguagem*, 16:1, 247-26. http://relin.letras.ufmg.br/revista/upload/11-Tommaso_Raso.pdf
Rocha, B. 2012. Características prosódicas do tópico em PE e o uso do pronome lembrete. MA Thesis. Universidade Federal de Minas Gerais. http://www.letras.ufmg.br/poslin/defesas/1466M.pdf
Rocha, B. 2013. Metodologia empírica para o estudo de ilocuções naturais do PB. Domínios de Lingu@gem 7(2), 109-148. http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/23747/13574
Scarano, A. (org.) 2003. *Macro-syntaxe et pragmatique: l'analyse linguistique de l'oral*. Roma: Bulzoni.
Searle, J. 1969. *Speech Act*. Cambridge: Cambridge University Press.
Searle, J. 1985. *Expression and Meaning*. Cambridge: Cambridge University Press.
Tucci, I. 2007. L'espressione lexicale dela modalità nel parlato spontaneo. PhD diss., Università degli Studi di Firenze.
Tucci, I. 2010. "Obiter dictum": la funzione informativa delle unità parentetiche. In M. Pettorino, A. Giannini & F.M. Dovetto (eds), *La comunicazione parlata 3*. Napoli: Università degli studi di Napoli l'Orientale, 635-654.
Vale, H. 2010. A Unidade Informacional de Parentético no Português do Brasil: uma análise baseada em corpus. MA Thesis. Universidade Federal de Minas Gerais. http://www.letras.ufmg.br/poslin/defesas/1350M.pdf
WinpitchW7. http://www.winpitch.com/ (accessed March 15, 2014).