Linguística de *Corpus* e C-ORAL-BRASIL: uma abordagem metodológica

Alessandra Emanuelle Macieira Silva

UFMG

Resumo: A Linguística de *Corpus* consiste em uma área da Linguística, que possibilita a sistematização, organização, coleta e identificação de dados de um determinado *corpus* linguístico. Por meio da Linguística de *Corpus*, é possível analisar determinada ocorrência textual e direcioná-la para o assunto de interesse utilizando-se, para tal fim, *softwares* que proporcionam o aprofundamento da análise. Tais programas computacionais permitem verificar a quantidade de ocorrências, formas verbais, concordâncias nominais e verbais, número de palavras, entre outras abordagens possíveis em um determinado *corpus*. Dessa forma, a proposta tem como objetivo a apresentação de conteúdo relacionado à Linguística de *Corpus*, visando retratar as possibilidades de estudo oferecidas por tal vertente da Linguística. Pretende-se destacar a importância e a eficiência do estudo de corpora feito pelo Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Faculdade de Letras na UFMG bem como da construção de corpora da fala espontânea pelo projeto C-ORAL-BRASIL.

Palavras-chave: Linguística de *Corpus*; *corpus*; C-ORAL-BRASIL.

SILVA, A.E.M. Linguística de *Corpus* e C-ORAL-BRASIL: uma abordagem metodológica. In: Jornada de Linguagens, Tecnologia e Ensino, 1, 2017. Timóteo. **Atas da […]**. Timóteo: CEFET-MG, 2017, p. 5-12. Disponível em: http://www.lite.cefetmg.br/publicacoes/publicacoes-da-1a-lite/. Acesso em: …

Linguística de *Corpus*: definição e aplicações

A Linguística de *Corpus* como área da Linguística contribui para o estudo de fatores linguísticos que seriam mais difíceis de compreender apenas por outros campos da Linguística. A associação entre a Linguística de *Corpus* e outras áreas da Linguística permite refinar o procedimento científico adotado para análise. A possibilidade de compilação de *corpora* (conjunto de *corpus*) de diversos tipos e análises linguísticas auxiliadas por meio de computadores ampliou as ferramentas que permitem o estudo da língua. Este artigo tem como objetivo apresentar em que consiste a Linguística de *Corpus* e sua aplicação em estudos de *corpora* de fala espontânea vinculados ao projeto C-ORAL-BRASIL, visando-se apresentar a Linguística de *Corpus* e sua intervenção em estudos linguísticos.

A compreensão da Linguística de *Corpus* pressupõe a definição do que é um *corpus*. Existem algumas definições distintas para determinados autores. Entretanto, a definição mais abrangente é apresentada por Sardinha:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador,

> com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (SARDINHA, 2004, p.18)

A descrição feita por Sardinha abrange vários pontos importantes que delimitam o que é um *corpus*. Primeiramente, o critério da origem dos dados consiste em sua autenticidade, isto é, ser composto por falantes nativos. O propósito do *corpus*, por conseguinte, indica que o mesmo deve ter a finalidade de ser um objeto de estudo linguístico. O conteúdo composicional do *corpus* também representa um critério que prioriza a naturalidade e autenticidade. Tendo-se delimitado que a Linguística de *Corpus* se utiliza de ferramentas computacionais, os dados devem apresentar formato legível por computadores além de serem representativos de uma língua ou variedade da língua. Por fim, a extensão do *corpus* constitui o último critério, que indica que o *corpus* deve ser vasto e com amplo número de textos.

O fato da Linguística de *Corpus* não delimitar o objeto de pesquisa e não determinar apenas um assunto — mas abranger vários temas dentro de uma mesma correspondência — contribui para não classificar tal área da Linguística como disciplina (SARDINHA, 2004). Em contrapartida, adotando-se a definição de metodologia como *um modo típico de aplicar um conjunto teórico* pode-se considerá-la como tal. Todavia, a definição de metodologia é variável e de acordo com as discrepâncias nas definições, a classificação da Linguística de *Corpus* também se altera. Desse modo, pode-se adotar outra categorização para a Linguística de Corpus: uma perspectiva.

De acordo com Hoey (1993) a Linguística de *Corpus* consiste em um meio para se atingir determinado objetivo, sendo que "Linguística de *Corpus* não é um ramo da linguística, mas a rota para a linguística". Desse modo, tal área compreende uma maneira para alcançar a linguagem e constitui-se com uma abordagem e não apenas como um instrumento (SARDINHA, 2004). Por meio das ferramentas e possibilidades de estudo proporcionadas através da Linguística de *Corpus* pode se chegar à linguagem e compreender seu uso e funcionamento, tal como é feito no projeto C-ORAL-BRASIL. Cabe ressaltar que o *corpus* C-ORAL-BRASIL é desenvolvido por pesquisadores do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Faculdade de Letras da UFMG e supervisionado e coordenado por Tommaso Raso e Heliana Mello, professores doutores titulares da mesma faculdade.

O uso de *corpora* computadorizados permite novos caminhos e possibilidades de estudos para os linguistas e questiona paradigmas estabelecidos por meio de abordagens empiristas. A Linguística de *Corpus* fundamenta-se em dois principais elementos conceituais: empirismo e visão probabilística da linguagem. Tais elementos são primordiais no estudo de *corpora* e representam a visão de Halliday (1991) em relação à linguagem. O primeiro consiste em um "quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema probabilístico" (SARDINHA, 2004). A conceituação de empirismo apoiada por Halliday (1991) estabelece, portanto, que os dados serão prioridade no estudo e posteriormente, a teorização. O segundo representa a visão da linguagem como probabilidade, isto é, a ocorrência dos traços linguísticos não apresenta a mesma frequência — alguns traços

podem ocorrer em quantidades distintas. Destarte, tal área da Linguística permite a compilação de *corpus* que possibilitam ao linguista análises da língua por meio dos elementos centrais definidos acima.

> A Linguística de Corpus ocupa-se da coleta e da exploração ele corpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. (SARDINHA, 2004, p. 3)

A definição de Sardinha evidencia a importância da aplicação do estudo de corpora em pesquisas linguísticas. A disponibilidade de *softwares* e programas desenvolvidos por estudos de Linguística de *Corpus* possibilita análises sintáticas, prosódicas e semânticas, o que por sua vez, amplia os recursos para desenvolver projetos e estudos além de otimizar o tempo e refinar e potencializar o conteúdo da pesquisa.

O projeto C-ORAL-BRASIL e a Linguística de *Corpus*

A área da Linguística de *Corpus*, apesar de valiosa e vantajosa não possui ampla aplicação e desenvolvimento no Brasil. A produção de estudos vinculados à Linguística de *Corpus* ainda é em formatos majoritariamente escritos e associados a gêneros acadêmicos (MELLO, 2012). Destarte, o desenvolvimento e aplicação da Linguística de *Corpus* são viabilizados pelo projeto C-ORAL-BRASIL, associado ao C-ORAL-ROM (CRESTI; MONEGLIA, 2005) o *corpus* europeu das quatro principais línguas românicas europeias. Por meio de recurso linguístico computadorizado, são realizados estudos teóricos e aplicados da fala espontânea e com base empírica (MELLO, 2012).

O *corpus* é composto por 139 textos e apresenta gravações de alta qualidade acústica divididas em contexto familiar e privado dos quais são subdivididos diálogos, monólogos e conversações. Ademais, os textos apresentados correspondentes aos áudios são alinhados no software *WinPitch*, que possibilita a visualização simultânea do texto e do espectrograma ao escutar o áudio (BOSSAGLIA, 2014).

O C-ORAL-BRASIL consiste em um *corpus* de fala espontânea do português brasileiro (PB) e representa a diatopia mineira, isto é, caracteriza a variação linguística de acordo com o lugar, sendo que majoritariamente, a composição do *corpus* é da fala metropolitana de Belo Horizonte. A definição de fala espontânea se faz necessária para a compreensão do desenvolvimento do projeto. Moneglia (2005) estabelece que fala espontânea corresponde a interações multimodais face a face, apresenta referência intersubjetiva a um espaço dêitico, programação simultânea à execução e comportamento linguístico contextualmente indeterminado (comportamento imprevisível). Desse modo, fala espontânea sugere um fluxo de fala natural e não programado. O fato de estar relacionado ao discurso natural e não planejado previamente, possibilita que através do estudo da fala espontânea seja possível analisar outras variações linguísticas: a alteração de acordo com o contexto comunicativo em que o falante está inserido (diafásica) e a variação de acordo com o estrato social e nível cultural do falante (diastrática).

A compilação do *corpus* C-ORAL-BRASIL possibilita estudos diversificados e relevantes para a Linguística. Por meio da análise da fala espontânea pode-se caracterizar as variações linguísticas, classificar a estrutura da fala espontânea e organização informacional, realizar estudos semânticos e morfossintáticos, dentre outros inúmeros estudos possíveis que abarcam a Linguística de *Corpus* (MELLO, 2012). Para a compilação do *corpus*, os dados são tratados por meio de principais etapas de desenvolvimento: gravação, transcrição dos dados e etiquetagem morfossintática. Por abrigar processos criteriosos para permitir a publicação final do *corpus*, há uma etapa de validação do *corpus* com o objetivo de verificar a consistência dos dados. Portanto, certifica-se que o *corpus* C-ORAL-BRASIL é consistente e válido, o que fornece às pesquisas e projetos desenvolvidos a partir de seu estudo, um alto grau de confiabilidade.

Corpus C-ORAL-BRASIL e fundamentação teórica

A arquitetura e segmentação do *corpus* obedece a uma fundamentação teórica baseada na Teoria da Língua em Ato (TLA), desenvolvida por Emanuela Cresti (1995, 2000a, 2000b). Trata-se de uma extensão da Teoria dos Atos de Fala de Austin (1962) e tem como base os *corpora* de fala espontânea do italiano (LABLITA). A TLA foi desenvolvida a partir de estudos de *corpora* durante quarenta anos, observando-se as regularidades e construções frequentes que possibilitam generalizações. Tais correspondências compõem o quadro teórico de uma teoria fortemente empirista por ser direcionada e condicionada pelo próprio *corpus* (RASO, 2012).

A Teoria da Língua em Ato tem como estruturação a análise pragmática da fala, assumindo-se que existam hierarquias distribuídas em níveis de acordo com a comunicação na fala. Logo, a distribuição dos níveis permite algumas individualizações que sempre são guiadas pela prosódia. A ilocução, portanto é considerada o primeiro nível de análise: o que o falante pretende ao produzir algum conteúdo locutivo, isto é, qual a sua intenção comunicativa: chamar, ordenar ou questionar, por exemplo. O segundo nível de análise consiste na estrutura informacional do enunciado, ou seja, em sua organização a partir de informações e divisões de acordo com as unidades presentes no enunciado. Somente após a individualização em unidades informacionais é possível atingir o último nível, que corresponde à análise sintática. Vejamos o exemplo que evidencia a importância da distribuição hierárquica dos enunciados retirados do *corpus* C-ORAL-BRASIL:

Exemplo 1 (link para arquivo de áudio)

*KAT: *o quê //*

*SIL: *copos // copos de Urano / que tem aí //*

*KAT: *copos de quê //*

*SIL: *Urano //*

*KAT: *Urano //*

*SIL: *é // Urano // Urano //*

A análise dos enunciados nos permite verificar alguns pontos principais: quatro enunciados são idênticos de acordo com a perspectiva semântica, morfossintática e ortográfica (*Urano*). Entretanto, tais enunciados não são idênticos considerando-se a prosódia. De modo simplificado, pode-se definir a prosódia como a emissão dos sons de fala. Uma das funções da prosódia é moldar "nossa enunciação imprimindo a "o que se fala" um "modo de falar" que é dirigido intencionalmente ou não ao ouvinte" (BARBOSA, 2012). Dessarte, a prosódia diferencia unidades idênticas sintaticamente e contribui para ressaltar a importância do estudo de *corpora* da fala espontânea. Ademais, é verificável a variação diastrática: a falante, possivelmente, não tinha conhecimento que os objetos em questão são os copos de Murano, fabricados na Itália, o que gera tal correspondência. As figuras abaixo foram retiradas do livro C-ORAL-BRASIL I (MELLO; RASO, 2012) e geradas pelo *software* de análise prosódica *WinPitch* que fornece a curva de entonação da fala de cada uma das participantes:

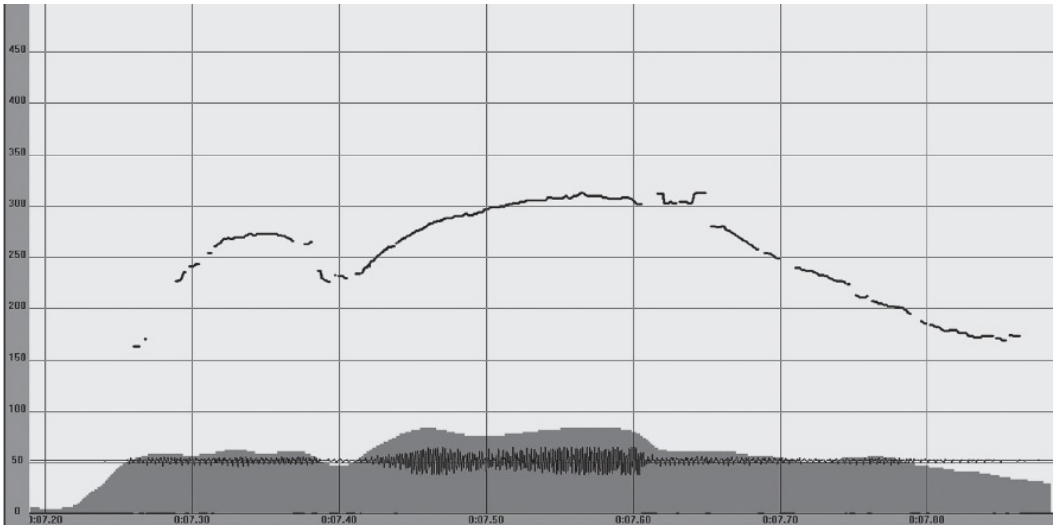

Figura 1: Ilocução de confirmação. Fonte: RASO, 2012, p. 96.

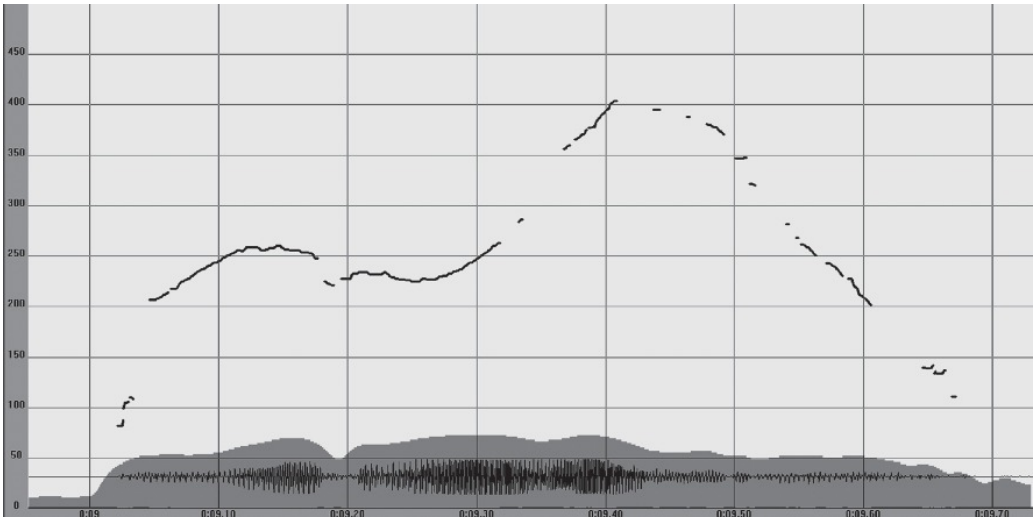

Figura 2: Ilocução de incredulidade. Fonte: RASO, 2012, p.97.

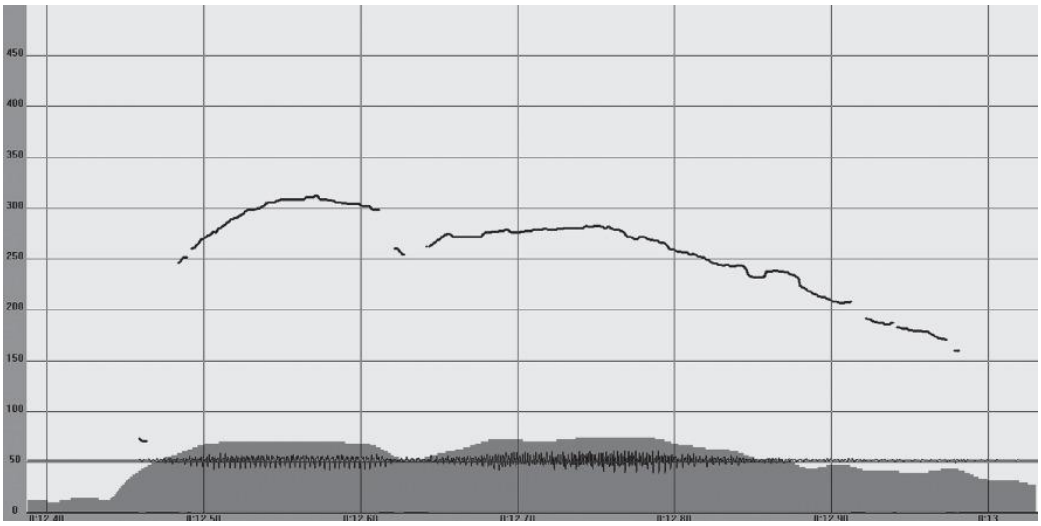

Figura 3: Ilocução de conclusão. Fonte: RASO, 2012, p. 97.

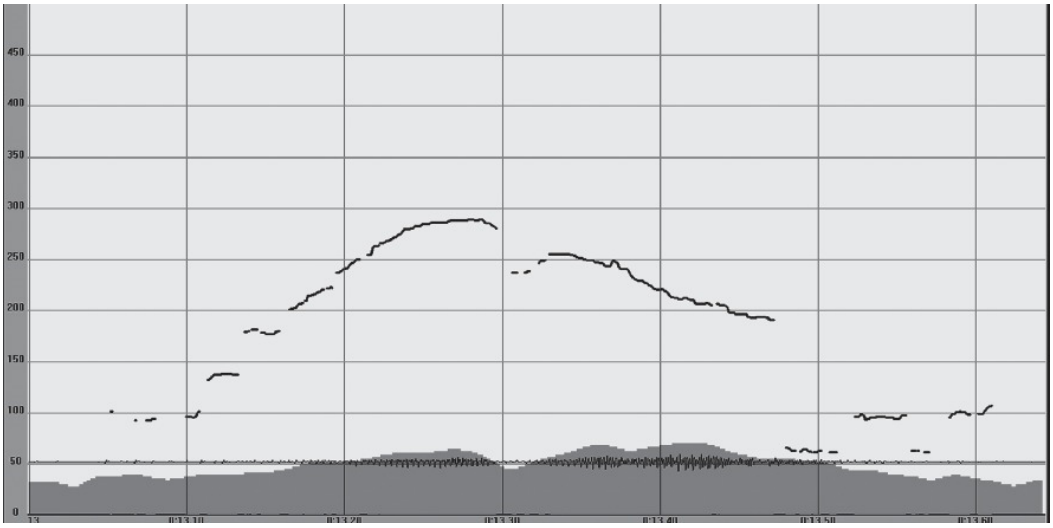

Figura 4: Ilocução de conclusão. Fonte: RASO, 2012, p. 98.

No exemplo acima, duas falantes produzem o mesmo conteúdo locutivo vinculando diferentes conteúdos ilocutivos perceptíveis apenas pela análise prosódica. O primeiro enunciado, que corresponde à falante *SIL* apresenta uma ilocução de confirmação (figura 1), ou seja, há uma afirmação após ser feito um questionamento prévio. A curva da figura 1 mostra que há uma intensidade parcial na pronúncia da falante, indicando uma confirmação. O enunciado posterior, por sua vez, expressa uma ilocução que indica incredulidade, na medida em que os dois últimos (*Urano // Urano)* expressam conclusões. A figura 2 destaca um pico na curva apresentada, o que indica que a falante está em dúvida sobre o que foi dito pela outra participante do diálogo. Apesar de as figuras 3 e 4 corresponderem à mesma ilocução, a curva

desta apresenta menor intensidade, o que denota uma mudança no estado de espírito do falante: a informação já foi dita e está sendo apenas repetida — mas não modifica o conteúdo ilocutivo (ainda corresponde a uma conclusão).

A caracterização das ilocuções presentes nos enunciados evidencia a importância da análise prosódica: apenas ao escutar os áudios é possível distinguir os enunciados, o que não é factível apenas pela escrita. Portanto, enunciados idênticos do ponto de vista semântico e sintático (mesmo conteúdo locutivo) apresentam conteúdos ilocutivos (intenções comunicativas) distintos e tal conteúdo ilocutivo é perceptível e vinculado pela prosódia (RASO, 2012).

A análise prosódica do enunciado permite verificar a importância da Teoria da Língua em Ato no estudo da fala espontânea. A inclusão do conteúdo prosódico na análise remete à fundamentação da teoria: "a característica mais inovadora da TLA é a inclusão do elemento prosódico na análise da fala espontânea" (BOSSAGLIA, 2014). Portanto, o conteúdo locutivo dos enunciados pode ser considerado independente com base nos critérios adotados pela TLA:

> O valor ilocutivo do enunciado é veiculado pela prosódia, que pode conferir autonomia pragmática a qualquer tipo de conteúdo locutivo, independentemente da sua autonomia semântica ou sintática. (BOSSAGLIA, 2014)

Os enunciados proferidos no exemplo 1 são, por conseguinte, independentes entre si devido ao fato de veicularem informações próprias em relação à prosódia. A independência pragmática de cada enunciado, isto é, o conteúdo de cada enunciado é autônomo em cada contexto situacional no qual o mesmo é veiculado. A observação da prosódia confere aos conteúdos referentes ao *corpus* de fala espontânea, características singulares que não podem ser compreendidas apenas pela sintaxe e semântica — apesar de tais áreas influenciarem o estudo.

Conclusões

A análise prosódica dos enunciados retirados do *corpus* C-ORAL-BRASIL evidencia a importância da aplicação da Linguística de *Corpus* em estudos de fala espontânea. Por meio de *softwares* associados a tal área de estudo, é possível verificar a construção comunicativa do falante no momento de sua fala, não se atendo apenas a critérios semânticos e morfossintáticos.

Os dados presentes no *corpus* C-ORAL-BRASIL denotam a relevância das variações linguísticas diafásica, diastrática e diatópica para a compreensão da comunicação na fala espontânea. Com base em enunciados produzidos pelos falantes, pode-se depreender e verificar como o contexto comunicativo, a camada social e nível cultural do falante e seu local de origem interferem na produção de enunciados, nas construções sintáticas e em conteúdos semânticos. Contudo, tal como é previsto na Teoria da Língua em Ato, a análise morfossintática e semântica é fundamentada, previamente, em uma análise prosódica da fala espontânea, sendo esta, o principal veículo comunicativo na fala.

A Linguística de *Corpus* é, portanto, uma área que proporciona caminhos de observações, estudos e pesquisas linguísticas com auxílio de meios computacionais que aperfeiçoam o trabalho e maximizam recursos. Por meio de programas e *softwares*, tal como o *WinPitch*, utilizado no *corpus* C-ORAL-BRASIL, é possível visualizar curvas de entonação na fala, verificar comportamentos característicos de falantes em relação à diatopia e diastratia, dentre outras consideráveis possibilidades. Os programas computacionais disponibilizam ferramentas que não competem ao trabalho manual e tornam a análise em questão mais refinada e fidedigna.

A associação da Linguística de *Corpus* ao trabalho realizado pelo projeto C-ORAL-BRASIL destaca a importância da aplicação de tal área em pesquisas que possuem relevância e credibilidade. A alta qualidade do *corpus* C-ORAL-BRASIL e a confiabilidade é resultado não apenas do trabalho humano excepcional e comprometimento da equipe, mas também da utilização de *softwares* e programas correspondentes à Linguística de *Corpus*. Desse modo, verificamos que a correspondência entre Linguística de *Corpus* e o projeto C-ORAL-BRASIL, fornece confiabilidade e credibilidade ao *corpus* e permite análises profundas e pertinentes em relação ao funcionamento da língua em uso, fundamentadas na Teoria da Língua em Ato.

Referências bibliográficas

BARBOSA, Plínio. Conhecendo melhor a prosódia: aspectos teóricos e metodológicos daquilo que molda nossa enunciação. *Rev. Est. Ling.*, Belo Horizonte, v. 20, n. 1, p. 11-27, jan./jun., 2012.

BOSSAGLIA, Giulia. Interface entre sintaxe e articulação informacional na fala espontânea: uma comparação baseada em corpus entre português brasileiro e italiano. *Caligrama*, Belo Horizonte, v. 19, n. 2, p. 35-60, 2014.

RASO, T; MELLO, H. (eds). *C-ORAL-BRASIL I*: Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG, 2012.

SARDINHA, Berber Tony. *Linguística de Corpus.* Barueri, SP: Editora Manole, 2004.