Apresentação

O *corpus* C-ORAL-BRASIL informal, os aparatos e o livro que o acompanham constituem o produto principal do projeto C-ORAL-BRASIL,[1] que vem sendo desenvolvido no Núcleo de Estudos em Linguagem, Cognição e Cultura (NELC) e no Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Faculdade de Letras da Universidade Federal de Minas Gerais desde 2007 sob a nossa coordenação. O projeto se inspira no projeto europeu C-ORAL-ROM[2] (Cresti; Moneglia, 2005) e na elaboração teórica que o grupo LABLITA[3] da Università degli Studi di Firenze tem desenvolvido em décadas de estudo sobre *corpora* de fala espontânea.

O livro e o DVD que o acompanha são organizados da seguinte maneira:

1. O livro:
   a) o primeiro capítulo, de autoria de Heliana Mello, coloca o *corpus* C-ORAL-BRASIL dentro do contexto brasileiro e dos *corpora* de fala do português brasileiro (PB) já produzidos;
   b) o segundo capítulo, de autoria de Tommaso Raso, apresenta o *corpus* e o compara com os *corpora* de espanhol, italiano, francês e português europeu do projeto europeu C-ORAL-ROM, no qual este projeto se inspira;
   c) o terceiro capítulo, também de autoria de Tommaso Raso, trata da teoria que inspirou tanto o projeto C-ORAL-ROM, e antes dele o *corpus* LABLITA (Cresti, 2000a), quanto este projeto, e

[1] <http://www.c-oral-brasil.org/>.

[2] <http://lablita.dit.unifi.it/coralrom/>.

[3] <http://lablita.dit.unifi.it/>.

que é importante para fundamentar a arquitetura e o sistema de segmentação do C-ORAL-BRASIL;

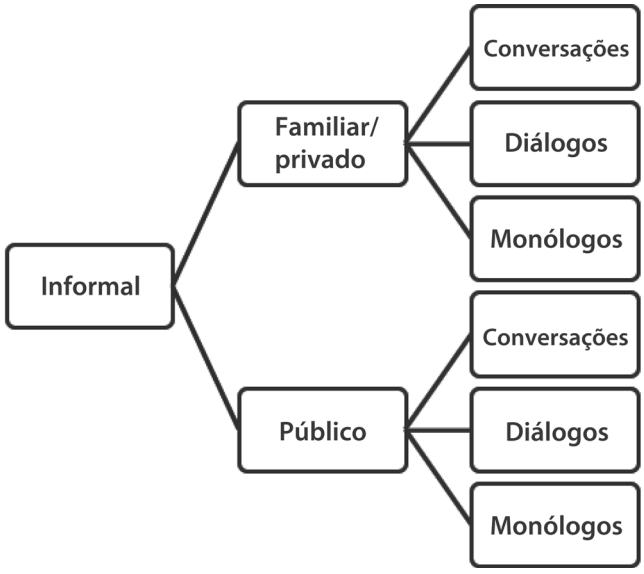
d) o quarto capítulo, escrito por Heliana Mello, Tommaso Raso, Maryualê M. Mittmann, Heloísa P. Vale e Priscila O. Côrtes, explica em detalhes o sistema de transcrição e de segmentação do *corpus* e mostra os resultados das validações relativas a essas duas operações;

e) o quinto capítulo, de Tommaso Raso e Maryualê M. Mittmann, fornece e discute algumas medidas da fala com base no *corpus* C-ORAL-BRASIL e nos *corpora* do projeto C-ORAL-ROM;

f) o sexto e último capítulo, de Eckhard Bick, apresenta o *parser* utilizado para a etiquetagem do *corpus* e seu desempenho.

2. O DVD:

a) a pasta *Multimedia Corpus* contém os arquivos de som, em formato wav, os arquivos de texto, em formato rtf, e os arquivos de alinhamento para o *software* WinPitch (www.winpitch.com), em formato xml; os arquivos, nomeados com base na ramificação de pertence e numerados em ordem progressiva, estão distribuídos dentro de pastas que agrupam as diferentes ramificações do *corpus*, segundo o seguinte esquema:



b) a pasta *Textual Corpus* contém todos os arquivos de texto em formato txt;

c) a pasta *PoS Tagged Corpus* contém 5 pastas:

- aquela dos arquivos etiquetados com o *parser* Palavras em formato txt;
- aquela dos arquivos etiquetados com o *parser* Palavras em formato xml;
- aquela dos arquivos etiquetados com uma versão simplificada do *parser* Palavras em formato txt;
- aquela dos arquivos etiquetados com uma versão simplificada do *parser* Palavras em formato xml;
- aquela com o arquivo com o *PoS Tagset*;

d) a pasta *Appendix* também contém 5 pastas:

- uma pasta contém as listas de frequência;
- uma pasta contém as planilhas com as medidas do *corpus* e as estatísticas relativas aos informantes;
- uma pasta contém os metadados de todos os textos do *corpus*;
- uma pasta contém a exemplificação de todos os tipos de quebras prosódicas marcadas nas transcrições;
- uma pasta contém as especificações do *corpus*;

e) a pasta *Book* contém o livro em pdf, com os exemplos de áudio linkados ao texto.

Este trabalho não teria sido possível sem a ajuda de muitas instituições e de muitas pessoas. Os organizadores agradecem à Fapemig, ao CNPq, à UFMG e em especial à Faculdade de Letras[4] pelo apoio financeiro e logístico. Um agradecimento especial vai para Emanuela Cresti e Massimo Moneglia, inspiradores e verdadeiros orientadores deste trabalho. Sem a paciência, o estímulo, a colaboração, a inteligência e a amizade de Emanuela e Massimo o *corpus* e este livro não existiriam.

[4] <http://www.letras.ufmg.br/>.

O trabalho deve muito também a uma grande quantidade de colaborações, dentro e fora da UFMG. Fora da UFMG, queremos agradecer a colaboração de Eckhard Bick e dos jovens pesquisadores do laboratório LABLITA: Alessandro Panunzi, Ida Tucci, Gloria Gagliardi e Lorenzo Gregori. Dentro da UFMG, agradecemos aos nossos alunos de pós-graduação e de iniciação científica e aos alunos da disciplina de Pragmática dos anos de 2008 e de 2009. Gostaríamos de explicitar os nomes de Maryualê M. Mittmann, Heloísa P. Vale, Priscila O. Côrtes, Bruna M. Rocha, Bruno Rocha, Bruno Alberto de Oliveira Mota, Adriana Ramos, Lucas L. Goulart, Andréia Ulisses, Luciano Alves de Deus, Flávia A. de Castro Leite, Janayna Carvalho, Cássia F. Oliveira, Raíssa C. Oliveira, Estefânia Melo, Elisa M. Franco, Renata Amaral. Agradecemos também o suporte técnico de Agnaldo Martins.

Tommaso Raso
Heliana Mello

As transcrições dos textos do *corpus*, além dos critérios não ortográficos explicitados no Capítulo 4, se baseiam no padrão ortográfico anterior à última reforma, já que esta não estava em vigor quando o projeto começou.