The C-ORAL-BRASIL corpus: methodological basis for the treatment of spontaneous speech

Maryualê M. Mittmann[1], Tommaso Raso[2], Heliana R. Mello[3]

[1]Faculdade de Letras – Programa de Pós-graduação em Estudos Linguísticos
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brazil

[2]Faculdade de Letras – Programa de Pós-graduação em Estudos Linguísticos
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brazil

[3]Faculdade de Letras – Programa de Pós-graduação em Estudos Linguísticos
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brazil

maryuale@ufmg.br, tommaso.raso@gmail.com, heliana.mello@gmail.com

Abstract. *This paper highlights the primary methods employed in the C-ORAL-BRASIL compiling process, i.e, recording, transcribing and segmenting oral texts. The C-ORAL-BRASIL is a Brazilian Portuguese corpus of spontaneous speech, designed for the study of informational structure. It is representative of the diaphasic variation, seeking to cover as many different comunicative situations as possible. This paper presents and exemplifies the processes of transcription and segmentation of speech into prosodic units as employed in our on-going research. It concludes with illustrations of some questions that the corpus will enable us to answer.*

1. Introduction

The C-ORAL-BRASIL is a spontaneous speech Brazilian Portuguese (BP) corpus, which was designed from its inception to study the informational structure of Brazilian Portuguese (BP) and its illocutions based on the Informational Patterning Theory [Cresti 2000]. It comprises the fifth branch of the C-ORAL-ROM [Cresti and Moneglia 2005], the reference corpora of the following four Romance European languages: Italian, Spanish, French and Portuguese.

The C-ORAL-BRASIL corpus constitutes a very important contribution to the registry of Brazilian Portuguese speech, since there are only very few corpora dedicated to oral language in Brazil. The corpus reliability and originality are ensured by the high acoustic quality recordings in natural contexts and the carefully monitored transcriptions, a process involving of prosodic segmentation marking, registering significant speech phenomena and text-to-speech alignment, wich is performed with the software WinPitch Pro [Martin 2004], a real time speech analysis tool.

2. C-ORAL-BRASIL general features

C-ORAL-BRASIL is being constructed with the same general architecture adopted by the C-ORAL-ROM project [Cresti and Moneglia 2005, Moneglia 2004], but with certain adaptations to the Brazilian sociolinguistic context. The entire corpus will be comprised of at least 30 hours of recordings, those being divided into 15 hours of informal speech and 15 hours of formal speech. The defining characteristics of the informal half of the corpus will be the focus of this paper, since the formal part of the C-ORAL-BRASIL constitutes a second stage of the project, with its own characteristics.

The 15 hours of informal recorded speech are distributed in a minimum of 100 texts made up of an average of 1500 words each. The level of variation which is focused upon in the C-ORAL-BRASIL corpus pertains to the diaphasic level, since this definitely has an impact on speech's informational structure variation. The diaphasic variation is represented in the corpus according to the following distribution, within informal speech: public (20%) versus family/private contexts (80%); within each context, three interactional typologies: monologue (1/3), dialogue and conversation (more than two participants) (2/3). Within each interactional typology, the goal is to maximize variation of communicative situations.

The majority of the texts are from one of the diatopic varieties of Mineiro speech [Zágari 2005], particularly the urban area of its capital city (Belo Horizonte). The corpus attempts to represent the diastratic variation to a certain extent, however, no statistical balance will be sought in this regard. In all the different constituent parts of the corpus, interactions among speakers from several socio-cultural strata are included.

3. Methodological steps in compiling the C-ORAL-BRASIL corpus

Since the C-ORAL-BRASIL is primarily designed for the study of informational structure and illocution in spontaneous speech, one of the major concerns regarding the compiling process of the corpus was to assure the registry of the largest possible diaphasic variation.

The monologic type corresponds to 30% of the corpus texts and encompasses topic variation (life narratives, interviews, work monologues, jokes, assorted narratives, etc.), variation as far as the recorded individual's profile is determined (family members, friends, clients, workmates, children, etc.), as well as variation in places where the recordings were carried (workplace, friends' and families' homes, restaurants, etc.). Dialogues and conversations, wich represent 70% of the texts, have the same above mentioned variables, but also present a broad variation as far as the activity being executed during the interactions.

3.1. Recordings and acoustic quality

The assurance of a large diaphasic variation in the corpus demands that the recordings are carried out in natural situations, and the interference of the researcher in the situation must be maintained to a minimum.

The recordings are realized with PDD60 Marantz digital recorders and Sennheiser Evolution EW100 G2 wireless kits (receiver, transmitter and clip-on microphone). This equipment supports recordings done in natural environments with a frequency rate of 22050 Hz, that allows the analysis of the intonation pattern, obtained from the speech signal through the fundamental frequency (F0) curve [Harrington and Cassidy 1999]. This acoustic quality also grants sufficient information for segmental analysis, ensuring vowel identification in the speech waveform.

3.2. Transcription

The transcriptions follow the CHILDES-CLAN system [MacWhinney 2000], implemented through the prosodic annotation criteria created to represent the speech segmentation in utterances and tone units [Moneglia and Cresti 1997], suited to the study of informational patterning [Cresti 2000]. The basis for transcribing is orthographic; still, several adaptations were established to better render some speech phenomena found in BP, like lexicalizations and grammaticalizations in progress. At the same time, the transcribed text must not present such complexity as to generate comprehension problems for the reader and excessive difficulties for the transcribers.

Space limits prevent a full discussion of transcription criteria, nevertheless, two relevant examples are be provided below. The slashes indicate intonational breaks perceived as terminal (//), delimiting utterances, and perceived as non-terminal (/) delimiting tone units.

a) Lack of person marking in several verb forms. Example:

DAN: o dia que eu mais Bola / da última vez que **nós foi** pro centro //

b) Subject pronoun cliticization in second and third person forms. você(s) > cê(s); ele > e'; ela > ea; eles > es; elas > eas. Example:

ROG: **ea** tá querendo andar um bocado / uai //

3.3. The segmentation training process

Segmenting speech into prosodic units (utterances and tone units) is done simultaneously during the transcription process, as both are based on acoustic perception. The linguists involved have completed a long training period in several stages. After the training period, the group of transcribers selected were submitted to tests, with the goal to obtaining homogeneity throughout transcriptions made by different people.

The potential transcribers were sorted into two groups, Group #1 and Group #2, according to the degree of aptitude in segmenting presented during the training period. The multi-rater Kappa statistic [Fleiss 1971] was used for assessing the reliability of agreement between the transcribers. The goal was to obtain an agreement greater than 80% concerning terminal breaks and greater than 60% for non-terminal breaks. Below a summarized description of the path followed by Group #1 is presented.

1. Segmentation of a dialogic text comprised of around 800 words and a monologic text comprised of 800 words. Kappa test results: 0.820 for the dialogic text and 0.750 for the monologic text.

2. Segmentation of a dialogic text comprised of 1500 words. Kappa test result: 0.839.

3. Segmentation of a monologic text comprised of around 1500 words. Kappa test result: 0.839.

4. Future Perspectives

We conclude this paper with one example study that the C-ORAL-BRASIL corpus allow us to pursue, which we find interesting for its large descriptive potential. The aim is to identify what belongs to speech characteristics and that which is specific of a given language/culture. We refer to the possibility provided by the C-ORAL-BRASIL for us to compare BP spontaneous speech to European Portuguese (EP) spontaneous speech, based not solely on segmental, morphosyntactic and lexical parameters, but also on those which pertain to the prosodic, informational and illocutionary domains.

Other uses of the corpus include its potencial as a database for technical fields such as automatic speech processing and synthesis. Since it provides a reliable database of segmented natural speech, it allows the understanding of the specific acoustic parameters used to signal different informational structures. That provides value information for refining speech recognition machines and developing more sophisticated tools for a more natural synthesis of speech.

References

Cresti, E. (2000), Corpus di italiano parlato, Accademia della Crusca, Firenze, 2 voll.

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

Fleiss, J. L. (1971), "Measuring nominal scale agreement among many raters", Psychological Bulletin, 76, 378-382.

Harrington, J. Cassidy, S. (1999), Techniques in Speech Acoustics, London: Kluwer Academic Publishers.

MacWhinney, B. J. (2000) The CHILDES project: tools for analyzing talk, Lawrence Erlbaum, Mahwah, 2 voll.

Martin, P. (2004) "WinPitch Corpus: A text to speech alignment tool for multimodal corpora", LREC, Lisbon, http://lablita.dit.unifi.it/coralrom/papers/index.html, July 2009.

Moneglia M. (2004) “Specifications of the C-ORAL-ROM corpus”, ELRA, Paris http://lablita.dit.unifi.it/coralrom/papers/Specifications-CORALROM.pdf, July 2009.

Moneglia M. Cresti, E. (1997) "Intonazione i criteri di trascrizione del parlato adulto e infantile". In: Bortolini, U.; Pizzuto, E. Il Progetto CHILDES Italia. Pisa: Del Cerro, pp. 57-90.

Zágari, M. R. L. (2005), "Os Falares Mineiros: Esboço de um atlas lingüístico de Minas Gerais", In: A geolingüística no Brasil: trilhas seguidas, caminhos a percorrer, Edited by V. A. Aguilera, Editora da Universidade Estadual de Londrina, Londrina, v. 1, p. 45-72.