The creation of a comparable minicorpus from the *Santa Barbara Corpus of Spoken American English*

INTRODUCTION AND OBJECTIVES

The present work consists in a report of the creation process of an American English minicorpus (AE minicorpus) from the *Santa Barbara Corpus of Spoken American English* – SBC – (DU BOIS *et al.*, 2000-2005). This work was carried out under the auspices of the C-ORAL-BRASIL project, which is devoted to the study of spoken language and the compilation of spontaneous speech corpora (RASO; MELLO, 2012), at the *Laboratório de Estudos Empíricos e Experimentais da Linguagem* (LEEL), at Federal University of Minas Gerais (UFMG), directed by Professors Tommaso Raso and Heliana Mello.

The AE minicorpus comprises 20 carefully selected texts to assure comparability to the minicorpora of the C-ORAL family (CRESTI; RASO, 2012) for Brazilian Portuguese (BP) and Italian (IT). The *Language into Act Theory* – L-AcT – (CRESTI, 2000) is the framework that provides the theoretical and methodological support for the creation of such minicorpora. It comprises a set of principles and methodologies for the empirical study of spontaneous speech, and is the result of decades study by Emanuela Cresti and other researchers at the LABLITA lab, University of Firenze.

Assuring its comparability to the BP and IT minicorpora was the main objective in the creation of the AE minicorpus. Thus, the overall design of the AE minicorpus is precisely that of those minicorpora, whose main features include:

1. diverse communicative situations;
2. monologues, dialogues and conversations proportionally represented;
3. transcription in CHAT format (MACWHINNEY, 2000);
4. prosodic segmentation;
5. informational tagging.

The minicorpus created expands the possibilities for cross-linguistic studies within the C-ORAL projects, providing researchers with data from a language of great academic reach. American English is now the first non-Romance language to be represented in a minicorpus specifically designed for studies within the L-AcT approach. Additionally, the AE minicorpus provides adequate means for the dissemination of the theoretical and methodological principles of the L-AcT approach.

The AE minicorpus will soon be available online to the academic community through the DB-IPIC platform (PANUNZI; GREGORI 2011), a queryable XML database designed to allow the study of linear relations among informational units in spoken language.

METHODOLOGY

The SBC is a body of 60 texts of spontaneous spoken American English recorded in many different locations within the United States. It contains a variety of people from different regional origins, ages, occupations, genders, and ethnic and social backgrounds. The corpus documents various ways in which people use language in their everyday lives, including telephone conversations, card games, food preparation, on-the-job talk, story-telling etc.

The choice of the SBC as the matrix corpus for the creation AE minicorpus was made based on its core characteristics. Firstly, the SBC is a spontaneous speech corpus, and that was an essential feature to allow comparability with the BP and IT minicorpora. Secondly, the diversity of communicative situations seemed to be compatible with the assortment found in the two pre-existing minicorpora. Lastly, the acoustic quality of the majority of the recordings is good enough to enable prosodic investigations. Additionally, the SBC is available online under a *Creative Commons license*.

The creation of the AE minicorpus followed a number of steps that can be divided in 5 phases as follows:

1. Text selection: search for recordings containing informal interactions, with good acoustic quality and varied communicative situations, as the BP and IT minicorpora feature;
2. Transcription: cleaning the original transcripts and implementing the transcription criteria and prosodic segmentation in accordance with the L-AcT methodology;
3. Alignment and first revision: text-to-sound alignment with the software *WinPitch* (MARTIN, 2005) and correction of eventual segmentation errors;
4. Second revision: the aligned texts were submitted to a second revision in order to discuss problematic cases regarding the prosodic segmentation;
5. Informational tagging: texts were manually tagged and information tags were assigned to each prosodic unit in accordance with the L-AcT principles. Texts were then submitted to another revision.

RESULTS

The AE minicorpus was prosodically segmented into tone units and utterances and aligned using the same methodological criteria adopted in the compilation of the C-ORAL minicorpora. The final product is the AE minicorpus, which comprises a total of 30,105 words in 20 texts. These numbers corresponds to the transcription of about 2 hours and 25 minutes of spontaneous speech recordings extracted from the SBC. The average text in the AE minicorpus has approximately 1,500 words, but there are three texts that count less than 1,000 words and two that count more than 2000 words.

The corpus is divided into familiar/private and public domains and contains 5 conversations, 8 dialogues and 7 monologues. Speech transcriptions were done in CHAT format (MACWHINNEY, 2000) with implementation of a prosodic annotation system (CRESTI; MONEGLIA, 1997).

Each of the 20 minicorpus texts comprises the following files:

a) Audio recordings in WAV format;

b) Transcriptions in TXT and RTF formats;

c) Header files with metadata of the recording in TXT format;

d) Text-to-sound alignment in xml format to be used with the software *WinPitch* (MARTIN, 2005).

REFERENCES

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

CRESTI, E.; RASO, T. *Annotation of information patterns according to Language into Act Theory in DB IPIC* - first release 2012, http://lablita.dit.unifi.it/app/dbipic/.

CRESTI, E.; MONEGLIA, M.; L´intonazione e i criteri di trascrizione Del parlato adulto e infantile. In: Bortolini, U. – Pizzuto, E. *Il Progetto CHILDES Italia*. Pisa: Del Cerro, 1997.

Du Bois, John W., *et al*. *Santa Barbara corpus of spoken American English*, Parts 1-4. Philadelphia: Linguistic Data Consortium, 2000-2005.

MACWHINNEY, B. J. *The CHILDES Project: tools for analyzing talk*. Mahwah: Lawrence Erlbaum Associates, 2000.

MARTIN, Ph. WinPitch Corpus: a text-to-speech analysis and alignment tool. In: Cresti, E.; Moneglia, M. (Eds.). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins, 2005

PANUNZI, A.; GREGORI, L. DB-IPIC: An XML database for the representation of information structure in spoken language. In: MELLO, H. R. *et al.* (Eds.). *Pragmatics and prosody: illocution, modality, attitude, information patterning and speech annotation.* Firenze: Firenze University Press, 2011, p. 133-150.