Oral corpora transcription and prosodic annotation validation

Heloísa VALE (UFMG)
Maryualê M. MITTMANN (UFMG)
Priscila CÔRTES (UFMG)

This paper concerns an innovative methodology for the transcription validation of oral corpora and the application of known methodology for the validation of prosodic segmentation. The validations are very important to test the limits of the statistic reliability of any study based on corpora. These methodologies were applied to the compilation of the C-ORAL-BRASIL corpus (Raso-Mello 2010), the fifth branch of the C-ORAL-ROM project (Cresti-Moneglia 2005). The Brazilian corpus is comprised of two halves: one formal and the other informal. The latter is already completed. The informal portion contains 139 texts and 210,000 words: 80% of family/private contexts and 20% of public contexts; 1/3 of dialogues, 1/3 of conversations and 1/3 of monologues.

The corpus is essentially based on the diatopy of the state of Minas Gerais, respects the diastratic variation, but has as its main goal the representation of the diaphasic variation, considered the major cause of the variation in speech structure. By means of using high quality wireless equipment, it was possible to record an ample variety of communicative situations, including those recorded during motion.

The transcriptions are orthographically based (MacWhinney 2000), but intend to document many features of speech, making therefore possible the study of phenomena in course of grammaticalization or lexicalization: cliticization of subject pronouns, loss of verbal morphology, loss of the verb to be in cleft structures, verbal serialization, apheresis and many others (Mello-Raso 2009). During the transcriptions the texts were prosodically segmented into utterances (pragmatically autonomous units with prosodic breaks perceived as conclusive) and tonal units (prosodic breaks perceived as non-conclusive within the utterance) (Moneglia-Cresti 1997). The transcribers/annotators went through a training process (Raso-Mittmann 2009).

Two validations were done concerning the prosodic segmentation: one after the beginning of the transcription of the recordings (after several discussion meetings and evaluations) and another after the whole corpus had been transcribed and revised for the first time, but before the consecutive revisions. The first validation consists of a significant methodological implementation: it establishes that the work begins only when there is expertise enough to guarantee a high quality standard. This way, the revision phases can indeed care for improvements in the transcription/segmentation. The validation consists of achieving a degree of agreement, measured by the Kappa test (Fleiss 1971), ≥ 0.8 for terminal breaks and ≥ 0.6 for non-terminal breaks, among 4 annotators. The first validation general results were: 0.84 (terminal breaks) and 0.66 (non-terminal), with some remarkable differences concerning dialogic and monologic texts. The second validation general results were: 0.86 (terminal) and 0.78 (non-terminal), with a radical reduction of the differences between dialogic and monologic texts. A qualitative study of the cases of disagreement was interesting to reveal not only prosodic aspects of BP that may induce uncertainty among the annotators, but also to acquire more expertise for future studies.

The validation of the transcriptions is new in the corpus linguistics methodological framework. The goal is to identify the degree of reliability of the transcriptions according to two perspectives: the first, more general, consists of the quantification of the percentage of utterances and of words wrongly transcribed; the second, essential when non-orthographic transcribing criteria are used, quantifies the reliability degree of each transcription criterion. Also here the validation has been developed in two parts: before the last revision and after the

conclusion of the corpus to be published. In each part a random sample of 10% of the utterances of each text was taken and examined in search of mistakes. The searching methodology was: two transcribers checked the transcriptions and, in case of disagreement, turned to a third transcriber (rare cases). Initially, we considered satisfactory a range of error not higher than 5% both for the total of words and for each phenomenon individually considered.

The results of the first phase (performed before the last revision of the whole corpus) have shown the presence of errors within 1.4% of the words of the sample. In the analysis of the errors concerning each criterion (35 phenomena, for instance você/ocê/cê or para/pra/pa) we have observed 3% of errors (37 errors over 1165 occurrences) concerning all phenomena. The second validation (performed after the last revision of the whole corpus) has taken into account 10% of the number of utterances of all 139 texts of the corpus. This sample consisting of 10%, 24,783 words and 3308 utterances received a different treatment. The percentage of every kind of errors (those concerning the transcription criteria and those concerning wrongly transcribed, missing or exceeding words) in the whole sample is 0.31%. On a subsample 5% the occurrence of all phenomena from the list of criteria as well as the frequency of orthographic forms have been the subject of analysis. 0.58% of errors concerning all criteria have been found (12 errors over 2119 words). Regarding the 5% left, only some criteria, whose frequency has been considered statistically insufficient, have been analyzed. The majority of the phenomena has shown a rate of less than 5% of errors, which guarantees the reliability of the transcriptions.

Another stage of the validation of the segmental level consisted on the analysis of the texts which constitute the minicorpus, a sample of the most representative and of best acoustic quality texts of the corpus. The minicorpus received morphological and informational annotations. During the informational annotation some segments were changed: some words were eliminated, others, added, and others yet, changed. Only 4% of the sample has gone through changes.

These validation methodologies represent a step forward in the search of reliability in oral corpora, adding reliabiblity to already known methodologies.

References

CRESTI, E.; MONEGLIA, M. (eds) (2005) C-Oral-Rom: Integrated Reference Corpora For Spoken Romance Languages. Amsterdam: John Benjamins.

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

MACWHINNEY, B. J. (2000) The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum, 2 vol.

MELLO, H.; RASO, T. (2009) Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. Revista Portuguesa de Humanidades, v. 13, p. 301-325.

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.

RASO, T.; MELLO, H. (2010) The C-ORAL-BRASIL corpus. In: MONEGLIA, M.; PANUNZI, A. (orgs.) Bootstrapping Information from Corpora in a Cross Linguistic Perspective. Firenze: Firenze University Press. p. 193-213.

RASO, T.; MITTMANN, M. M. (2009) Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. Revista de Estudos da Linguagem, 17, 73-91.