Lexical category distribution in a spontaneous speech corpus of Brazilian Portuguese

This paper discusses the distribution of lexical categories in the C-ORAL-BRASIL corpus (Raso & Mello, 2012), an informal spontaneous speech corpus of Brazilian Portuguese. The corpus is comparable in architecture and segmentation criteria with the four C-ORAL-ROM corpora (Cresti & Moneglia, 2005). C-ORAL-BRASIL presents 75% of private/familiar texts and 25% of public texts; for each context 1/3 of texts are monologues, 1/3 dialogues and 1/3 conversations. Each individual text is sized at about 1,500 words. The corpus is text to speech aligned through the WinPitch software. Its main goal is to document diaphasic variation with the widest range of different communicative situations. Segmentation of the speech flow is done through prosodic criteria resulting in utterances and tone units. Utterances, defined as the smallest pragmatically interpretable unit, end with a prosodic break perceived as conclusive, while tone units – which make up utterances - end with a non-terminal break. Transcriptions were made with quasi-orthographic criteria that aimed at representing grammaticalization and lexicalization phenomena in speech, while also attempting to maintain easy readability of the texts and consistency in transcribers’ perceptions.

C-ORAL-BRASIL was entirely POS tagged and grammatically marked using the PALAVRAS parser (Bick, 2012), especially adapted to deal with the particularities of this corpus. Therefore, keeping consistency in relation to the segmentation of the speech flow, the analytical unit for POS and grammatical tagging was each tone unit within a given utterance. Tone unit boundaries, marked through prosodic breaks, were read as punctuation by the parser. Hence, a semicolon was adopted as the equivalent to the (/) terminal breaks (alternating with ‘...’ for interruptions), and a comma for the non-terminal breaks (/) The tagging and parsing show that the morphological and syntactic organization of speech is quite different from that of written texts since the latter is organized around clauses and sentences, while the former relies on information units and utterances that may or may not hold a predication, therefore making meaning production depended not only on contextual clues but on prosodic patterns as well. Additionally, not all information units behave syntactically in the same way since utterances are made up of both textual and dialogical units. The former carry the burden of providing the core meanings codified syntactically and semantically by the illocution and represented by the utterance content. The later, on the other hand, are responsible for keeping the communication channels open between the interlocutors and usually are impoverished as far as lexical content is concerned.

Previous work analysing the informational organization of the three different textual types in C-ORAL-BRASIL, that is, monologues, dialogues and conversations, along with the grouping in the familiar/private versus public environments, has shown that monologues tend to be richer in complex textual units, followed by dialogues and conversations. This points out to the straightforward hypothesis that the more informationally elaborate a text type is, the heavier it will be on lexical content, since the bulk of semantic meaning in language in general is portrayed by Nouns, Adjectives, Adverbs and Verbs.

In order to test this hypothesis, we have extracted all the words POS tagged as Nouns, Adjectives, Adverbs and Verbs from the corpus, correlating the score of lexical categories with the interaction type by using an R script. The resulting clustering can be seen in the dendrogram portrayed in figure 1 below.
As predicted, the dendrogram shows that clades lead to the clustering of similar textual types, whereby the major groupings coincide with conversations as one branch of a clade while dialogues and monologues are gathered together in the other branch. Consistently with the initial hypothesis, dialogues and monologues are separate branches of the same clade. There is only a small portion of texts that behave anomalously (lower branch in first level clade). These (4 dialogues and 1 monologue, out of 139 texts) require qualitative inspection in order to be best described.

The quantitative study reported in this paper has demonstrated that there is a strong association between textual type and lexical category density in spoken Brazilian Portuguese. Monologues are the textual type that exhibit the most elaborate and extensive textual information unit organization. This is also the case for its lexical category organization. Following down on an information structuring complexity scale, there come dialogues and lastly, conversations. Coherently, the quantitative results demonstrate that as far as lexical category distribution is concerned, dialogues average closer to monologues than they do to conversations, supporting the view that conversions are more oriented towards pragmatic organization than morphosyntax.

References
