The C-ORAL-BRASIL I: Reference Corpus for Spoken Brazilian Portuguese

Tommaso Raso, Heliana Mello, Maryualê M. Mittmann

Universidade Federal de Minas Gerais
Av. Antonio Carlos, 6627, Belo Horizonte, Brazil
tommaso.raso@gmail.com, heliana.mello@gmail.com, maryuale@gmail.com

Abstract

C-ORAL-BRASIL I is a Brazilian Portuguese spontaneous speech corpus compiled following the same architecture adopted by the C-ORAL-ROM resource. The main goal is the documentation of the diaphasic and diastratic variations in Brazilian Portuguese. The diatopic variety represented is that of the metropolitan area of Belo Horizonte, capital city of Minas Gerais. Even though it was not a primary goal, a nice balance was achieved in terms of speakers' diastratic features (sex, age and school level). The corpus is entirely dedicated to informal spontaneous speech and comprises 139 informal speech texts, 208,130 words and 21:08:52 hours of recording, distributed into family/private (80%) and public (20%) contexts. The LR includes audio files, transcripts in text format and text-to-speech alignment (accessible with WinPitch Pro software). C-ORAL-BRASIL I also provides transcripts with Part-of-Speech annotation implemented through the parser system Palavras. Transcripts were validated regarding the proper application of transcription criteria and also for the annotation of prosodic boundaries. Some quantitative features of C-ORAL-BRASIL I in comparison with the informal C-ORAL-ROM are reported.

Keywords: C-ORAL-BRASIL, Brazilian Portuguese, spontaneous speech

1. Introduction

The language resource presented in this paper is the product of the C-ORAL-BRASIL Project, which is associated with the Laboratory of Empirical and Experimental Language Studies (LEEL) based at Minas Gerais' Federal University (UFMG). The main goal of the project is to offer a spontaneous speech corpus of Brazilian Portuguese for the study not only of lexis and morphosyntax, but also of pragmatic categories, such as information structure and illocution.

C-ORAL-BRASIL I (Raso & Mello, 2012; Raso & Mello, 2010; Raso & Mello, 2009) consists of the informal [1] branch of the C-ORAL-BRASIL Project and represents the diatopic variety of Minas Gerais state, mainly from the metropolitan region of the capital, Belo Horizonte. The corpus was built to be comparable to the C-ORAL-ROM resource (Cresti & Moneglia, 2005) by adopting the same architecture, transcription format, segmentation criteria and text-to-speech alignment tool and methods.

The primary aim of this language resource (LR) is the documentation of diaphasic variation in spontaneous speech which is taken to be the major reason for the structural variation in speech. In order to be considered spontaneous, speech events must not accomplish a pre-existing text, neither in part nor in whole (Nencioni, 1983). Spontaneous speech occurs in multi-modal face to face interactions in which there are: an inter-subjective reference to a deitic space; concurrent mental programming and vocal execution; unpredictable linguistic behaviour (Moneglia, 2005).

[1] The formal branch is currently under construction and will comprise formal communicative situations as well as media and telephone situations.

2. Technical Information

C-ORAL-BRASIL I is a multimedia LR that makes available for the user:

- Audio recordings (wav files);
- Orthographic transcription complemented with prosodic annotation (txt and rtf files);
- Text-to-speech synchronization through WinPitch Pro software (Martin, 2004) (xml files);
- Metadata for each recording session (txt files);
- Orthographic transcription with Part of Speech annotation performed by the parser system Palavras (Bick, 2000; 2012) (txt and xml files).

Recorded sessions stored in "wav" files (Windows PCM, 22.050 Hz, 16 bit) were carried out with a Marantz PMD660 Professional Solid State Recorder and high resolution, non-invasive wireless equipment, mostly mono-directional clip-on microphones (Sennheiser EK/SK 100 G3). Whenever there were more than two interactants, an analog mixer (Behringer XEXYX 1222 FX) was used. In a few occasions, an omnidirectional microphone (Sennheiser MD 421-II 4) was used.

This equipment ensured a high acoustic quality which is, in the majority of cases, sufficient for F0 calculation through WinPitch, even though several recordings took place in rowdy contexts with background noise.

60% of the recordings have high or extremely high acoustic quality (A and AB), while only 23% have low acoustic quality (C). Inevitably, the low quality is more common in multiparty conversations that, due to their nature, have more voice overlapping and present more challenges regarding microphones. Acoustic quality for each corpus audio file is provided in the metadata and follows the classification detailed in Table 1.

Tag	Description
A	Extremely high quality. Almost no voice overlapping and/or background noise. Trustable F0 computation for (practically) the entire file.
AB	High quality. Low voice overlapping and/or background noise. Trustable F0 computation for (practically) the entire file.
B	Medium quality. Some voice overlapping and/or background noise. Trustable F0 computation for most part of the file.
BC	Mid low quality. Some voice overlapping and/or background noise. Trustable F0 computation for at least 60% of the file. Audio is clear for listening throughout the entire file.
C	Low quality. Some voice overlapping and/or background noise. Trustable F0 computation for at least 60% of the file. Some portions of the audio may not be clear for listening.

Table 1: Description of acoustic quality tags.

Low quality sessions were used only if a particular aspect of interest was also present. Note that some interesting daily situations necessarily come with background noise. As examples from the corpus we may cite: purchase in shops or supermarket, soccer playing, party and several others.

Table 2 shows the total number of recorded sessions in each corpus node with the correspondent acoustic quality.

Corpus node	A	AB	B	BC	C	Total
Family/private conversations	8	11	4	6	5	34
Public conversations	1	2	0	1	5	9
Family/private dialogues	7	14	6	5	3	35
Public dialogues	5	2	2	1	1	11
Family/private monologues	13	10	10	1	2	36
Public monologues	3	4	4	2	1	14
Total	40	43	25	14	18	139

Table 2: Acoustic quality of audio files.

3. Design

C-ORAL-BRASIL I comprises 21 hours, 8 minutes and 52 seconds of speech recordings, which corresponds to a total of 208,130 transcribed words in 139 text files. The mean word number per text is 1,500. Only 10 texts are larger than 2,000 words (with a maximum 4,800 words) and 16 some are smaller than 1,000. However, they all keep textual autonomy.

Transcriptions follow the CHAT format (MacWhinney, 2000), with implementation of prosodic boundary annotation (Moneglia & Cresti, 1997). Details regarding transcription and prosodic annotation criteria are given in section 4.

The corpus is made up of family/private context (159,364 words e 105 texts) and public context (48,766 words and 34 texts). In each of the two contexts the number of texts was equally dived among monologues, 2-person dialogues and multiparty conversations.

In monologues the structure of speech depends mainly on textual typology: life history, professional explanation, argumentative text, joke, recipe, fable, etc. In the dialogues and conversations the variation is basically linked to the activity that the interlocutors are carrying: a conversation among friends at home will be structured in a very different way from a row between a couple, or from a dialogue between a shoe store attendant and a costumer, or from an interaction among football teammates in a game, etc. It is clear that the illocutions that should be performed through speech change radically. Each situation stimulates the emergence of different speech acts, different turn sizes, different utterance size and structure, larger or smaller silence periods, etc.

The diastratic variation is represented in the 362 speakers recorded in the corpus. Sex, age, origin, and schooling are registered for 68.23% of speakers. The nearly 30% who were not documented for social parameters consist of speakers who entered the recording context unpredictably. As far as number of words uttered, undocumented speakers make up 1.91% of the corpus. On the other hand, such a large number of unpredicted speakers in the corpus supports the fact that recordings were not scripted or controlled and were, in fact, spontaneous.

The female/male balance is very precise as far as number of uttered words is concerned: 50.36% of words are uttered by (203) females and 49.64% of words are uttered by (159) males.

Likewise, there is a balance regarding number of uttered words/age rate, see Figure 1 below.

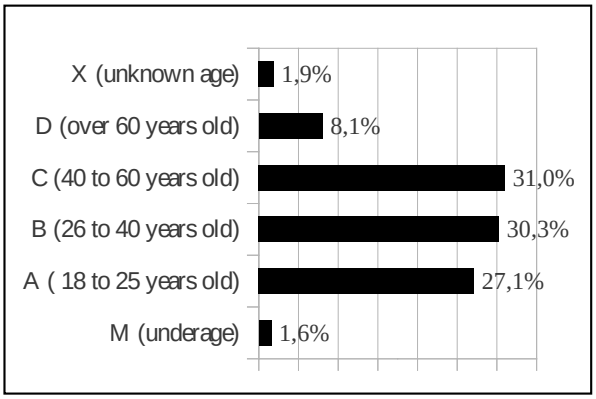

Figure 1: Percentage of uttered words per age group

Schooling is divided into 3 groups: level 1, speakers without formal schooling or that have up to 7 years of schooling (incomplete basic level); level 2, speakers up to undergraduate degree as long as not having a profession related to university degree; level 3, speakers who work in professions dependent on a university degree. It also shows a good balance, as shown in Figure 2.

Diastraty is, therefore, very well balanced in all aspects, favouring speakers who belong to middle to middle-high schooling level, which allows for the corpus to be representative of a synchronous standard. Nevertheless, lower schooling levels are also represented.

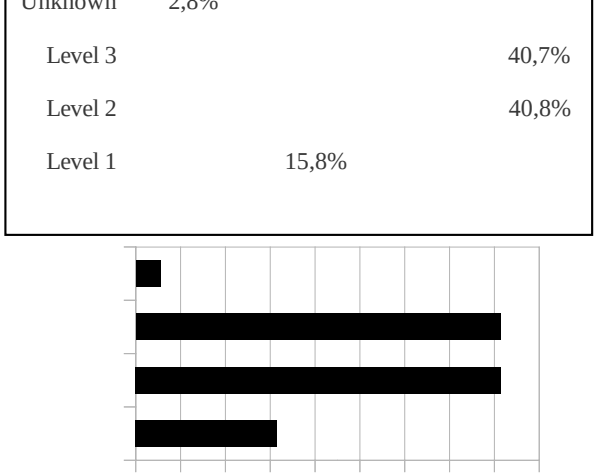

Figure 2: Percentage of uttered words per schooling level

Although speakers' occupations are much diversified, a significant percentage works in activities related to education (students and teachers in all schooling levels and areas, as well as school heads, school administrative staff, etc).

A corpus with these dimensions cannot include the diatopic variation. Therefore, the chosen diatopic variety was that of Belo Horizonte (this is the same procedure adopted for the C-ORAL-ROM in which the chosen cities were Florence, Aix-Marseille, Madrid and Lisbon). Speakers origins are the following: 138 are from Belo Horizonte, 89 from other municipalities in Minas Gerais (many belonging to the Belo Horizonte metropolitan area), 19 from other Brazilian states, 2 from other countries and 119 without their origin being documented but which represent an insignificant portion of word number in the corpus.

4. Transcriptions

The process of transcribing Brazilian speech involved the development of specific criteria for the representation of spontaneous speech phenomena, the training of transcribers for the annotation of prosodic boundaries, as well as a series of revisions and the content validation.

Speech transcriptions were done in accordance with the CHILDES-CLAN system (MacWhinney, 2000) with implementation of a prosodic boundary annotation system developed by Moneglia and Cresti (1997).

The transcriptions of the C-ORAL-BRASIL have an orthographic basis, but several adaptations in the notation system were introduced. Transcriptions attempt to capture phenomena that reflect lexicalization and grammaticalization in progress. These are phenomena such as: nominative pronouns cliticization, verbal paradigm reductions, demonstrative reductions, absence of verb ser (to be) in clefts and other focus structures, aphaeresis, among others.

Nevertheless, there is a necessity for balancing the rendering of linguistic phenomena and the readability of the text or the feasibility of the transcription. The transcription criteria cannot impose excessive difficulties for the transcribers, especially in those cases in which the perceptibility of the phenomenon to be transcribed is such as to render improbable a high level of agreement among transcribers. Besides, the resulting transcript cannot generate comprehension problems for the reader.

We provide below a small sample that illustrates the transcription system adopted. Asterisks indicate opening of dialogic turn; capital letters signal the speaker; slashes signal prosodic boundaries (see section 4.1) and angled brackets indicate voice overlapping.

Excerpt from bfamdl01:

**FLA: <brigada / moça> //*
thanks lady
**REN: <tá> // tá certo // <brigada> //*
ok all right thanks
**MDS: <brigada> //*
thanks
**REN: vamo lá //*
lets go
**FLA: aonde nós temo que ir //*
where must we go

In this example we find some words transcribed according to non-orthographic criteria. In the first three turns we have the aphaeresis of the words *obrigada* (*brigada*) and *está* (*tá*). In the two final turns we see the deletion of the final /s/ from both main verbs: *vamos* (*vamo*) and *temos* (*temo*).

The adoption of non-orthographic criteria for speech transcription raised the awareness of linguistic phenomena that had not received proper attention in linguistic studies done so far on Brazilian Portuguese.

This new approach to speech transcription seeks to reveal aspects of the systemic evolution of Brazilian Portuguese, focusing especially on the variety of Minas Gerais.

Thus, the C-ORAL-BRASIL allows the examination of the extent to which this diatopic variety could anticipate changes morphosyntactic phenomena that would extend to other linguistic varieties of Brazil.

4.1 Prosodic boundaries annotation scheme

In C-ORAL-BRASIL I, just as in C-ORAL-ROM, the speech flow is segmented according to prosodic criteria. The segmentation (prosodic boundary annotation) is based on the Language Into Act Theory – LAcT (Cresti, 2000), which assigns the utterance as the reference unit to speech.

The utterance is defined as the smallest prosodically and pragmatically autonomous speech unit, individualized through a prosodic boundary perceived as concluded (terminal break) and represented in transcription as a double slash (//). The utterance constitutes the linguistic counterpart of a unit of action; a locution that corresponds to an illocution (Austin, 1962; Moneglia, 2011).

Excerpt from bpubdl04:

**ELI: ih //*
(interjection)
**MUR: nũ entra não //*
it does'nt fit
**ELI: não //*
no

The utterance may also be prosodically segmented into smaller units, by means of prosodic boundaries that do not signal the completion of an autonomous speech unit (non-terminal breaks). Non-terminal breaks are represented with a single slash (/). These segment the utterance into tone units, which correspond to information units.
Excerpt from bfamdl09:

**FLA: eu achava que o Picasso era mais / velho //*
I thought that Picasso was older
**LUC: mais velho / tipo / de quando //*
older like from which time

Two other break types are represented in transcriptions. The first is a terminal break that marks utterances which were interrupted either by the speaker's will or by outside factors. It is represented through the symbol (+).
Excerpt from bfamcv12:

**GIL: mas isso / &he / complicado no sentido assim +*
but that &he is complicated in the sense that
**CAR: é &total + é muito mais leve do que um telhado //*
it's &total it's much lighter than a roof

The second type signals word retracting, which are represented by a slash and a number within brackets ([/n]). The number refers to the number of words retracted by the speaker.
Excerpt from bfammn03:

**ALO: não / mas aí [/1] é [/1] aí / é coisa séria //*
no but that is that is something serious

4.2 Transcribers

The first step consisted of training the transcribers. This training process had two goals: (i) to enable the transcribers to identify intonation cues that characterize different prosodic boundaries in the speech flow and differentiate them from other acoustic information that does not involve the segmentation of speech into prosodic units; (ii) to ensure the highest possible degree of coherence and consistency in the annotation of prosodic boundaries throughout the corpus.
The transcribers' team was formed by undergraduate and graduate students linked to C-ORAL-BRASIL Project. The transcribers went through a process of academic and methodological training that enabled them to acquire the theoretical basis and the procedures adopted in the corpus for the speech transcription and segmentation into prosodic units.
Transcribers' performance and skill levels were evaluated during training. The team was subdivided into two groups: Group 1 (G1) was formed by 3 expert and highly skilled transcribers. Group 2 (G2) was formed by 4 expert transcribers moderately skilled.
This division was aimed at allowing better control on the transcription and review processes. Thereby better performance transcribers were responsible for reviewing the work of less skilled transcribers.
The methodological training process involved alternate sessions of segmentation tasks, inter-rater agreement testing and feedback. Each group of transcribers first annotated the prosodic boundaries of the same text individually and in isolation from each other. Next, a inter-rater agreement test measured the group's agreement on the annotation of prosodic boundaries. The results were then evaluated and discussion sessions were held within each group.
This process was repeated until each group had an inter-rater agreement score considered sufficient for the beginning of transcriptions work. This consists of a significant methodological implementation, because it establishes that speech transcription work only begins when there is enough expertise to guarantee high quality standards.

5. Validation

C-ORALBRASIL I underwent two validations: one regarding the annotation of prosodic boundaries and other concerning the transcripts segmental content. Both were done internally and the evaluators were expert transcribers.

5.1 Validation of prosodic boundary annotation

There are multiple simultaneous prosodic cues involved in the perception of prosodic boundaries, such as pitch reset, fall of intensity, pause, rhythm, final lengthening and initial rush. Therefore speech segmentation cannot be performed through acoustic data alone, and even perceptual judgments sometimes are not entirely coherent (Moneglia et al., 2010). Since the annotation of prosodic breaks is done during speech transcription based only on perception, it is important to establish a validation process that ensures the consistency of such annotation.
The methodology applied in the validation of the C-ORAL-BRASIL prosodic segmentation involved a pre-validation and a final validation. The pre-validation occurred before the beginning of the transcription process and corresponds to the degree of expertise obtained by transcribers by the end of the training period.
The final validation occurred when the entire corpus was transcribed, but before the final revisions were performed. Final validation was accessed only in G1, since this group was the only responsible for the final revisions.
Kappa statistics (Fleiss, 1971) was used to assess agreement between annotators in each group (G1 and G2). The goal was to obtain an agreement greater than 0.8 for terminal breaks and greater than 0.6 for non-terminal breaks in final validation (reference values established by C-ORAL-ROM standards, see Danieli et al., 2004 and Moneglia et al., 2005). The task consisted in hearing and segmenting transcribed texts (deprived from any prosodic annotation) into utterances, signalling the perception of terminal and non-terminal breaks. Each annotator worked autonomously and without any consultation.
Table 3 presents the results for the pre-validation in groups 1 and 2.

Agreement type	Group 1		Group 2	
	dial	mon	dial	mon
General agreement	0.78	0.76	0.77	0.82
Terminal breaks	0.87	0.71	0.85	0.83
Non-terminal breaks	0.58	0.66	0.66	0.75
Break absence	0.84	0.86	0.81	0.87

Table 3: Pre-validation of prosodic boundary annotation.

Scores for Group 1 were obtained after 3 sessions of training. Scores for Group 2 were obtained after 5 training sessions in the case of dialogues and only after 8 training sessions in the case of monologues. Differences observed between the two text typologies are related with intrinsic difficulties regarding, particularly, the differentiation between terminal and non-terminal prosodic breaks in some monologues and are much related to the speaker's characteristic intonation.
Table 4 shows details regarding the percentage agreement reached by the two annotators groups, revealing the different skill degrees between the two.

Agreement type	Group 1		Group 2	
	dial	mon	dial	mon
Total agreement	85	83	79	87
Terminal breaks	13	6	12	6
Non-terminal break	7	13	12	8
Break absence	65	65	55	73
Partial agreement	15	16	19	12
Terminal vs non terminal break	5.2	7.5	5	5
Non-terminal break vs break absence	9.9	8.7	14	7
Total disagreement	0.3	0.5	2.4	1.5
Terminal break vs break absence	0.1	0.2	0.7	0.7
Terminal breaks vs non-terminal break vs break absence	0.1	0.2	1.6	0.8

Table 4: Percentage of agreement/disagreement in the annotation of prosodic boundaries.

Total agreement refers to the percentage of cases where all transcribers annotated the same type of prosodic boundary (terminal, non-terminal or absence of prosodic boundary). Partial agreement refers to the percentage of cases where (i) at least one of the transcribers signalled the presence of a terminal prosodic break and at least one of the others signalled a non-terminal break in the same position; or (ii) at least one of the transcribers signalled the presence of a non-terminal prosodic break and at least one of the others signalled break absence in the same position. Finally, total disagreement refers to the percentage of cases where (i) at least one of the transcribers signalled the presence of a terminal prosodic break and at least one of the others signalled absence of break in the same position; or (ii) each one of the transcribers signalled a different value for the same position.
The higher expertise of G1 is attested by the lower percentage of total disagreements.
The final validation results are shown in Table 5.

Agreement type	Group 1		
	overall	dial	mon
General agreement	0.86	0.86	0.85
Terminal breaks	0.87	0.87	0.86
Non-terminal breaks	0.78	0.78	0.78
Break absence	0.91	0.91	0.90

Table 5: Final validation of prosodic boundary annotation.

The increase in inter-annotator agreement is clear, especially regarding non-terminal breaks. Agreement on terminal breaks went from 0.58 and 0.66 in dialogues and monologues respectively to 0.78 for both.
Additionally there are no score differences between text typologies (monologues and dialogues) indicating that transcribers improve their skills during the transcribing work.

5.2 Validation of transcripts

Two validations of transcriptions took place. The first was carried out before the last revision of transcriptions began. It comprised a sample of 5% of utterances randomly extracted from each corpus text (7,484 words). The goal of this first validation was to access the percentage of errors and, more importantly, to verify if the non-orthographic criteria was implemented successfully with an acceptable margin of error. The results obtained in this first validation would also orient the final revision of transcripts.
The sample was examined by 2 expert transcribers who searched for overall errors and incorrect application of non-orthographic criteria. Overall errors include misspellings and typos, word deletions (absence of a word in the transcript that is present in the audio source) and word insertions (presence of a word in the transcript that is lacking in the audio source).
The first validation results were positive and indicated that the larger proportion of errors is due to improper application of non-orthographic transcription criteria.
Table 6 shows the results for the first validation of the transcripts.

Error type	Errors/words	%
All errors	140/7,484	1.9
Overall errors	104/6,319	1.6
Incorrect spelling	45/6,319	0.7
Word insertion	19/6,319	0.3
Word deletion	40/6,319	0.6
Misapplication of transcription criteria	37/1,165	3.2

Table 6: First validation of transcripts.

The second validation was conducted after the last corpus revision, when all transcripts were ready to be published. A new random sample of 5% of the utterances from each text was checked (8,243 words) by one expert transcriber. The percentage of errors should not exceed 5% of words. The results for the final evaluation are presented in Table 7 and show that the last revision indeed eliminated some of the errors, especially regarding the transcription of words that should be written according to the non-orthographic criteria defined in the specifications (0.6% of errors).
Misspellings, word deletions and insertions decrease from of 1.6% to 0.9%. The total percentage of error in the corpus is about 0.81%.

Error type	Errors/words	%
All errors	67/8,243	0.8
Overall errors	55/6,124	0.9
Incorrect spelling	23/6,124	0.4
Word insertion	19/6,124	0.3
Word deletion	13/6,124	0.2
Misapplication of transcription criteria	12/2,119	0.6

Table 7: Final validation of transcripts.

The final validation indicates correctness of 98.9% to 99.3% of words (95% confidence interval). The results attest the transcripts high accuracy regarding the application of orthographic as well as non-orthographic criteria.

6. Spontaneous speech features for Brazilian Portuguese

In this section we present some statistics from the C-ORAL-BRASIL I corpus related to the natural reference units of spontaneous speech, that are dialogic turns and utterances. We show how these vary in size and complexity according to the corpus branch.
We also compare some data from C-ORAL-BRASIL I with the informal section of C-ORAL-ROM.

6.1 Dialogic turn

The first natural unit of reference of a spoken text is the dialogic turn, defined as a continuous stretch of speech from the same speaker, delimited by the speech of another one. Table 8 shows the average number of utterances per turn and the average number of words per turn.

Interaction type	Utterances / Turn			Words / Turn		
	min	max	mean	min	max	mean
Conversations	1.2	2.1	1.5	4.4	14	7.4
Dialogues	1.5	3.5	1.8	6.4	25.2	9.6
Monologues	1.9	90	3.0	12.8	44.9	28.6

Table 8: Mean values for utterances per turn and words per turn.

Minimum value (min) refers to the text corpus C-ORAL-BRASIL with the lowest mean and the maximum value (max) refers to the text with the highest mean.
Mean values of utterances per turn listed in Table 8 show that the structure of turns varies greatly among monologues. In fact, it is very difficult to find a perfect exemplar of monologue in spontaneous speech. The reason for this is that in spontaneous speech the interlocutor will always interact with the speaker. When this happens, she often does it by manifesting her agreement with short and structurally simple utterances.
In general, the number of utterances per turn is a good measure of the texts level of interactivity. Usually, the higher is the number of utterances per turn, the smaller is the degree of interactivity and the higher is degree of textual elaboration. The word per turn rate also corroborates this observation.

6.2 Structural complexity of utterances in informal spontaneous speech

In C-ORAL-BRASIL I, like in C-ORAL-ROM, the reference unit for speech is the utterance, as defined by Cresti (2000). Utterances can have a simple or a compound structure, depending on whether they present internal prosodic segmentation or not. Simple utterances are constituted by one single tone unit and compound utterances are constituted by two or more tone units.
According to Cresti (2005), the choice of a simple or compound structure is connected to the structure of the communicative event (dialogic or monologic). The greater or lesser structural complexity of utterances is related to a greater or lesser textual elaboration.
Table 9 shows the percentage of simple (constituted by one single tone unit) and compound (constituted by two or more tone units) utterances in Brazilian Portuguese (BP) in comparison to the languages represented in C-ORAL-ROM: European Portuguese (EP), Italian (IT), Spanish (SP) and French (FR). Regarding their internal structure, conversations and dialogues behave in the same way, so they are considered together in the category "dialogic".

Corpus	Dialogic		Monologic	
	simple	compound	simple	compound
BP	58.7	41.3	43.2	56.8
EP*	50.2	49.8	32.4	67.6
IT*	52	48.5	30.5	69.5
SP*	57.8	42.2	32.4	67.6
FR*	69.2	30.8	44.1	55.9

*Source: Cresti, 2005, p. 222.

Table 9: Percentage of simple and compound utterances in C-ORAL-BRASIL I and informal C-ORAL-ROM

The type of interaction (monologic or dialogic) seems to represent a significant variable in the structure of utterances. Monologues have a greater structural complexity than dialogic interactions in all languages.

In the BP, in dialogues and multi-party conversations (dialogical) simple utterances represent 58.7%, while compound utterances make up to 41.3%. In Monologues, complex utterances (56.8%) are more common than simple utterances (43.2%).
PE has a similar distribution of simple (50.2%) and compound (49.8%) utterances in dialogic interactions. These data attest some structural differences regarding the recording sessions of dialogic interactions in EP.
Italian (IT) is the language which has the highest proportion of complex utterances in monologues (69.5%), while French is the language with the highest proportion of simple utterances in dialogic interactions (69.2%).
The data presented here provide some inter-linguistic evidence that the type of interaction between participants in a communicative situation gives rise to different linguistic structures in speech.

6.3 Standard measurements for Romance Languages

In this section we present the standard measurements in the domain of Romance Languages laid out by the C-ORAL-ROM Project and reported by Moneglia (2004) and by Cresti and Moneglia (2005) in the C-ORAL-ROM resource, incorporating Brazilian Portuguese data.
As stated by Moneglia (2004), specific standard variation parameters can offer an important measurement of spoken language variability. Such measurements help to determine language-dependent and language-independent reference values.
The parameters are:

- Mid-Length of Utterances in words (MLU);
- Mid-Length of the dialogic turn in words (MLTw);
- Speed in words per second (Speed w);
- Mid length of the tone unit in words (MLTone).

Table 10 shows the results for the 5 Romance Languages from C-ORAL informal corpora.

Parameter	BP	EP*	IT*	SP*	FR*
MLU	6.16	7.54	6.51	7.81	14.49
MLTw	11.37	22.92	13.22	16.93	26.16
Speed w	2.76	3.08	2.56	3.19	3.48
MLTone	3.37	2.86	2.71	3.10	4.97

*Source: C-ORAL-ROM DVD (Cresti; Moneglia, 2005).

Table 10: Standard variation parameters for spontaneous speech in C-ORAL-BRASIL I and informal C-ORAL-ROM

Brazilian Portuguese follows the same general tendencies registered for C-ORAL-ROM languages. Mid-Length of Utterances (MLU) in informal speech is predictable in almost all languages (BP, EP, IT and SP), while French records a higher average of words per utterance.
The mid-Length of the dialogic turn indicates the level of interactivity among speakers in a communicative situation. Highly interactive situations tend to lead to greater dialogic turn shifts between speakers. The average number of words per turn in BP is the lowest among the compared corpora. In this item, BP is closer to Italian than to EP.
Speed in informal speech seems to be a more or less constant parameter among all five Romance languages (around 3 words per second).
In theory, we expect that the mid-length of the tone unit (MLTone) varies in accordance with the intonation, rhythmic properties and syllabic structure of each language. In C-ORAL-BRASIL one can notice a large number of long tonal units. The ability to produce long tonal units (considering the number of words), is probably due to the fact that the speech represented in the C-ORAL-BRASIL seems to be a more stress timed variety, as opposed to Italian, which is a syllable-timed language and, coherently, has the smallest number of words per tone unit.
Moneglia (2004) proposes that the high values of MLTone in French may be explained by the word weight in terms of number of syllables, since speech syllabic reduction of words with respect to the graphic representation in French is systematic.

7. Conclusion

The C-ORAL-BRASIL corpus is the first Brazilian Portuguese spontaneous speech aligned corpus. It is based on the C-ORAL-ROM LR, implementing several of its methodological aspects.
The chosen diatopic variety is that of the metropolitan Belo Horizonte area, in the state of Minas Gerais. Its dimensions are about 35% bigger than those of each individual corpus in the C-ORAL-ROM.
The diaphasic variation in C-ORAL-BRASIL I is much larger than that of the informal Italian C-ORAL-ROM sub-corpus, which is the most varied within that project. This was possible due to the careful planning of recording contexts; the selection of the best one third of the overall number of recordings made and to the very modern recording equipment used which allowed for excellent quality recordings in noisy natural environments as well as recordings with subjects in motion. Subjects are always carrying some action other than speech in most recordings.
An innovative transcription methodology for the study of language change in Brazilian Portuguese was implemented. Besides that, new validation methodologies for the transcription and segmentation validations were developed with a view to reach optimal agreement levels before the actual segmentation process was started, so as to guarantee that after the revision process was concluded, the results would be highly reliable.
The measures related to the size and composition of turns, utterances and tone units are important indicators of the degree of interactivity among speakers in the recorded sessions. It also allows the comparison of speech parameters between informal speech of all five Romance Languages represented in the C-ORAL resources.

8. Acknowledgements

The C-ORAL-BRASIL project was funded by the National Council for Scientific and Technological Development (CNPq), the Foundation for Research Support of Minas Gerais (Fapemig), the Faculty of Letters of the Federal University of Minas Gerais (Fale/UFMG) and by Santander Bank.

We thank the coordinators of the C-ORAL-ROM project Massimo Moneglia and Emanuela Cresti for the constant help and support, as well as the members of the Linguistic Laboratory of the Italianistic Department of the University of Florence (LABLITA): Alessandro Panunzi, Ida Tucci, Lorenzo Gregori and Gloria Gagliardi.

9. References

Austin, J. (1962). *How to do things with words*. Oxford, UK: Oxford University Press.

Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus: Aarhus University Press.

Bick, E. (2012). A anotação gramatical do C-ORAL-Brasil. In: T. Raso & H. Mello (2012). (Eds.), *O C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, pp 223- -254.

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

Cresti, E. (2005). Notes on lexical strategy, structural strategies and surface clause indexes in the C-ORAL-ROM spoken corpora. In: M. Moneglia & E. Cresti (Eds.), *C-ORAL-ROM: integrated reference corpora for Spoken Romance Languages*. Amsterdam-Philadelphia: John Benjamins, pp. 209 - -256.

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

Danieli, M. et al. (2004). Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM." In *Prococeedings of the 4th LREC Conference*. Paris: ELRA.

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

Martin, Ph. (2004). *WinPitch Corpus: A text to Speech Alignment Tool for Multimodal Corpora*. Available at: http://www.winpitch.com.

MacWhinney, B. J. (2000). *The CHILDES Project. Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum.

Mello, H., Raso, T. (2009). Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades*, 13(1), pp. 153 --178.

Mello, H.; Raso, T.; Mittmann, M.; Vale, H.; Côrtes, P. (2012) Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In T. Raso, & H. Mello (Eds.), *O C-ORAL-BRASIL I: um corpus de referência da fala espontânea do português brasileiro*. Belo Horizonte: Editora UFMG, pp. 125 --176.

Moneglia, M. (2004). Measurements of Spoken Language Variability in a Multilingual Corpus: Predictable Aspects. In: *Prococeedings of the 4th LREC Conference*. Paris: ELRA.

Moneglia, M. (2005) The C-ORAL-ROM resource. In: M. Moneglia & E. Cresti (Eds.), *C-ORAL-ROM: integrated reference corpora for Spoken Romance Languages*. Amsterdam-Philadelphia: John Benjamins, pp. 01 - -70.

Moneglia, M. (2011). Spoken Corpora and Pragmatics. *Revista Brasileira de Linguística Aplicada*, 11(2) pp 479 - -519.

Moneglia, M., Cresti, E. (1997). L´intonazione e i criteri di trascrizione del parlato adulto e infantile. In: U. Bortolini & E. Pizzuto (Eds.), *Il Progetto CHILDES-Italia: Contributi di ricerca sulla lingua italiana*. Pisa: Del Cerro, pp. 57 - -90.

Moneglia, M. et al. (2005). Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. In: E. Cresti & M. Moneglia (Eds.), *C-ORAL-ROM: integrated reference corpora for Spoken Romance Languages*. Amsterdam-Philadelphia: John Benjamins, pp. 257 - -276.

Moneglia, M. et al., (2010). Challenging the Perceptual Relevance of Prosodic Breaks in Multilingual Spontaneous Speech Corpora: C-ORAL-BRASIL/C-ORAL-ROM. In: *Proceedings of Speech Prosody 2010 Satellite Workshop - Prosodic Prominence Perceptual and Automatic Identification.* Chicago: Université de Neuchâtel.

Nencioni, G. (1983). *Di scritto e di parlato: Discorsi Linguistici.* Bologna: Zanichelli,.

Raso, T., Mello, H. (2009). Parâmetros de compilação de um corpus oral: o caso do C-ORAL-BRASIL. *Veredas*, 13(2), pp. 20 - -35.

Raso, Tommaso and Mello, Heliana. (2010). The C-ORAL-BRASIL corpus. In: M. Moneglia & A. Panunzi (Eds.) *Bootstrapping Information from Corpora in a Cross Linguistic Perspective*. Firenze: Firenze University Press, pp. 193 –213.

Raso, T. & Mello, H. (2012). (Eds.) *O C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG.

Raso, T., Mittmann, M. (2009). Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem*, 17(2) pp. 73 - -91.