THE C-ORAL-BRASIL PROJECT FOR BRAZILIAN PORTUGUESE SPOKEN CORPORA

EL PROJECTE C-ORAL-BRASIL PER A CORPUS ORALS DEL PORTUGUÈS BRASILER

Lúcia de Almeida Ferrari & Giulia Bossaglia
Universidade Federal de Minas Gerais
ferrari.lu@gmail.com / giulia.bossaglia@gmail.com

Abstract: In this paper we present a specific subset of spoken corpora of the C-ORAL family, namely the C-ORAL-BRASIL corpora of spontaneous Brazilian Portuguese (BP). Stemmed as the non-European branch of the C-ORAL-ROM project (Cresti & Moneglia 2005), the C-ORAL-BRASIL project has compiled third generation corpora of spoken BP, outstanding not only as specific BP corpora, but also as a model tool for the study of spoken language in general, also thanks to some methodological and technological improvements. Beside the resources for the study of spoken BP, a set of minicorpora compiled for specific studies on information structure (also in languages other than BP) are presented, together with other ongoing compilation processes developed within the C-ORAL-BRASIL research group. All the published resources are available for download at <www.c-oral-brasil.org>.

Key words: C-ORAL-BRASIL project; spoken corpora; compilation best practices; Brazilian Portuguese

Resum: En aquest article presentem un subconjunt de corpus orals de la família C-ORAL, concretament el corpus C-ORAL-BRASIL de portuguès brasiler espontani (PB). Derivat de la branca no-europea del projecte C-ORAL-ROM (Cresti & Moneglia 2005), el projecte C-ORAL-BRASIL ha aplegat uns corpus orals de PB de tercera generació, el qual destaca no sols com a corpus de PB, sinó també com una bona eina per a l'estudi de la llengua parlada en general, gràcies a algunes millores metodològiques i tecnològiques. A més dels recursos per a l'estudi del PB oral, presentem un conjunt de minicorpus creats per a l'estudi específic de l'estructura informativa (també en altres llengües a més del PB); així mateix, també tractem altres processos de compilació que estem desenvolupant actualment en el grup de recerca C-ORAL-BRASIL. Tots els recursos publicats estan disponibles a <www.c-oral-brasil.org> i es poden descarregar.

Caplletra 69 (Tardor, 2020), p. 201-220. ISSN 0214-8188, ISSN versió electrònica 2386-7159
Rebut: 14/10/2019. Acceptat: 15/02/2020. DOI: 10.7203/Caplletra.69.17269

Paraules clau: projecte C-ORAL-BRASIL; corpus orals; recollida de dades i bones pràctiques; portuguès brasiler

1. INTRODUCTION

In this paper we present a specific subset of spoken corpora of the C-ORAL family, namely the C-ORAL-BRASIL spoken corpora of Brazilian Portuguese (from now on, BP). Stemmed as the non-European branch of the C-ORAL-ROM project (Cresti & Moneglia 2005), the C-ORAL-BRASIL project has compiled third generation corpora of spoken BP, outstanding not only as specific BP corpora, but also as a model tool for the study of spoken language in general.

In the following subsections, we draw attention on some methodological issues that are relevant in the design of a spoken corpus, and we present a brief overview on the state of the art of BP spoken corpora.

In section 2, we present in detail the C-ORAL corpora specifications and methodological practices, focussing on some improvements achieved by the C-ORAL-BRASIL project, while section 3 is dedicated to a brief overview on some specific resources conceived as tools for studies specifically focussed on information structure —the C-ORAL-BRASIL informationally tagged minicorpora. In section 4, we present some final remarks.

1.1 THE DEMANDS OF STUDYING SPOKEN LANGUAGE

Within the last few decades, corpus linguistics has proved itself to be extremely valuable for linguistic research in several areas, and corpora can nowadays be considered as one of the most relevant tools for empirical studies.

For what regards speech studies, for a long time they have been conducted mainly on transcripts (i.e. *written* texts) of actual spoken interactions (see Mello 2014 for a more detailed state of the art of spoken corpora).

Within different approaches, more or less complex transcription systems (since Jefferson 1984) have been used to render the richness of typically spoken phenomena, such as hesitations, self-corrections, intensity increases and decreases, interruptions, paralinguistic sounds, and so on —but of course a lot of information is lost when spoken language is analysed without paying the due attention to the prosodic level.

For these reasons, third generation spoken corpora must comply with specific requirements demanded by the complexity of speech. First, they cannot fail to include the audio files of the recorded sessions, because prosody fulfils so many functions in speech, which are not fully transferable on the transcript, even by very rich and complex notation systems (which can result in less readability, at that).

Nonetheless, the availability of the signal is not enough *per se*: text-to-sound alignment improves significantly the way the researcher accesses the audio files. Through specific alignment softwares (e.g. WinPitch, ELAN, Praat, among others) it is possible to access simultaneously the transcript and the audio file (and, often, also the spectrograph) or to easily recover a specific portion of the recorded session, instead of searching for it throughout a non-aligned audio file.

A second main issue in the design of a spoken corpus concerns its transcription criteria and the way they adequately incorporate linguistic information —prosodic in the first place. The speech continuum is organized in smaller intonation units, signalled by different kinds of prosodic breaks corresponding to a combination of varied phonetic cues (intensity, duration, f0 movements, syllabic alignment, lengthening, among others, see Barth-Weingarten 2016 for a thorough review), provided with perceptual prominence. Thus, terminal (i.e. perceived as conclusive) and non-terminal (i.e. perceived as non-conclusive) breaks segment the flow of speech, with crucial consequences on the kind of linguistic relation between contiguous words. Consider the following sequence of words in BP:

não	*é*	*pra cima*
NEG	be.3SG	above

Since BP is a pro-drop language, there are at least two possible segmentations (hence, interpretations) of the sequence: *não é pra cima.* «it's not above» with no prosodic break between the words but the final one (here signalled by the full stop) vs. *não, é pra cima.* «no, it's above», with a non-terminal break after the negation (signalled by the comma) plus the final one. Thus, the presence or absence of the prosodic break will determine if the scope of the negation is the verb *é* or not —semantic and syntactic relations between words within the speech flow are strictly dependent on how this is organized into intonation units through prosodic cues. It follows that it is of the utmost importance that a spoken corpus includes the notation of this kind of prosodic information in its transcripts, since it is the prosodic segmentation that determines the kind of linguistic relations existing between words.

Other requirements in the compilation of a spoken corpus depend, of course, on the compilers' aims, i.e. on the variety they want their corpus to be representative of. The compilation of a *spontaneous*, i.e. non-planned, speech corpus demands a lot in terms of design. Real world interactions comprise different genres, registers, numbers of participants, human activities, and so on —clearly, the recording process itself is much more complicated than simply putting a recorder on a table while a group of people chat. In the compilation of the C-ORAL-BRASIL corpora, the recordings aimed at the greatest possible degree of diaphasic variation: a wide variety of communicative situations (Informal, Telephonic, and Formal in Natural Context corpora) or genres (Media corpus) was recorded, as we describe with more detail in 2.1 and 2.2. In this way, a significant diversity of speech acts and, hence, sociolinguistic features of the participants were documented, also thanks to specific methodological improvements in the recording process itself (cf. 2.3).

It is clear, then, that compiling a spoken corpus requires not only a lot of planning effort, but also the action of an articulated team of researchers: the great variety of recorded situations was possible only through hard work by every team member in finding useful contacts to have access to the many informal and formal spontaneous situations, in collecting the sociolinguistic information for each recorded participant together with the signed consent form required by the new ethical norms for scientific research, and so on. Not all recording processes end successfully: technical problems of the recording device, or complications in the consent form collecting (speakers changing their mind about giving consent to the use of the recording session; unavailability of some speakers, especially in what concerns the recording of formal situations, and so on), or quick changes in the team composition, a big deal of whom is often made up by young undergraduate students, and many other real-life contingencies make the recording (and, more broadly, compiling) effort greater. Compiling a spoken corpus means recording much more than what will actually be part of the final product.

1.2 Brazilian Portuguese spoken corpora: a concise overview

Before detailing the C-ORAL-BRASIL corpora specifications, a brief overview on the main resources for the study of spoken BP is suitable, even to understand the C-ORAL-BRASIL's novelty.

Back in 2004, the *Banco do Português*, compiled by the *Pontifícia Universidade Católica de São Paulo* (see Sardinha 2004: 9), reunited more than 200 million words of written and spoken texts. The corpus is not of public access, and the spoken sec-

tion comprises only transcripts, i.e. written texts, without their (aligned) audio files. The same holds for another important, but completely public, corpus for Portuguese varieties, the *Corpus do Português* (Davies & Ferreira 2006; <https://www.corpusdoportugues.org>). This huge (45 million words) historical corpus of European and Brazilian Portuguese has been enlarged since 2016 by a new subsection (1 billion words) of written texts from webpages from Portugal, Brazil, Angola and Mozambique. Beside the preponderance of the written section over the spoken one, it is worth noticing the absence of the audio files, on one hand, and the fact that the oral section is made up only by a single, not completely spontaneous, genre, that is, interviews from different tv and radio shows, collected from their websites.

One of the major projects for the compilation of spoken corpora for BP is certainly the NURC - *Projeto Norma Linguística Urbana Culta*, associated to different research units at different Federal Universities across Brazil. The NURC-RJ from the *Universidade Federal do Rio de Janeiro* (<http://www.nurcrj.letras.ufrj.br>) was the first nucleus of the project back in the 70s. The aim of the project is to collect data of spoken standard BP, today from different diatopic varieties across the country. Not all the data of the various NURC project research units were published, and the most part of the published ones include, one more time, only transcripts. The NURC-RJ makes available only a sample of texts provided with the audio files, but without text-to-speech alignment. The sources for the spoken texts are formal situations such as academic classes, conferences, and so on, plus researcher/informant interviews on different topics, and dialogic interactions between informants recorded by a researcher —then, not completely spontaneous situations.

The *Corpus de Referência do Português Contemporâneo* (<http://www.clul.ulisboa.pt/pt/component/content/article?id=714:crpc-corpus-de-referencia-do-portugues-contemporaneo>) compiled by the *Centro de Linguística da Universidade de Lisboa* comprises, between the European and other Portuguese varieties, a BP section (approximately 3 million words) including spoken texts. They come from informant/researcher interviews only, and, besides the transcripts, the audio files and the text-to-speech alignment (through the *EXMARaLDA* software: Schmidt 2004) are made available. While it is of course noteworthy that the spoken subsection displays some of the mandatory features for a really adequate spoken corpus, its size remains very small: approximately 100 thousand words for European Portuguese (downloadable from the website), and 86 spoken interactions from other varieties, published in four CD-ROMs available for purchase.

Other Brazilian projects for the compilation of spoken corpora are the *Corpus do Português Brasileiro Contemporâneo* compiled by the *Universdiade do Estado de São*

Paulo Araraquara, and the *Português Falado do Ceará* compiled by the Universidade Federal do Ceará, but they are not public yet.

From this brief overview (a more detailed state of the art of BP spoken corpora can be found in Mello 2012) emerges that the compilation of BP spoken corpora in Brazil is still incipient, compared to other resources available for other languages. In this sense, as it will be shown in the next sections, the C-ORAL-BRASIL corpora represent ground-breaking resources for the study of spoken BP.

2. THE C-ORAL-BRASIL PROJECT

The C-ORAL-ROM project (Cresti & Moneglia 2005) has published four comparable spoken corpora for the main European Romance languages, each one compiled by a specific team: the Italian corpus (compiled by the LABLITA lab at the Florence University: <https://www.letterefilosofia.unifi.it/vp-309-lablita.html>), the French corpus (compiled by the DELIC project at the *Université de Provence*: <http://www.elda.org/en/proj/coral/corpus_delic.html>), the Spanish corpus (compiled by the Linguistics lab at the *Universidad Autónoma de Madrid*) and the European Portuguese corpus (compiled by the *Centro de Linguística da Universidade de Lisboa*: <http://www.clul.ulisboa.pt/en/resources-en>).

The C-ORAL-BRASIL project stemmed as the non-European branch of the C-ORAL family, aiming at compiling spoken BP corpora and other speech resources. Its headquarters are found at the LEEL lab - *Laboratório de Estudos Empíricos e Experimentais da Linguagem* (<http://www.letras.ufmg.br/leel>) at the University of Minas Gerais (Belo Horizonte, Brazil). So far, the compiled resources are divided into C-ORAL-BRASIL I (Informal corpus; Raso & Mello 2012) and C-ORAL-BRASIL II (Formal, Telephonic, Media corpora; Raso *et al.* forthcoming).

All C-ORAL corpora were compiled based on the specific theoretical and methodological tenets of the *Language into Act Theory* (L-AcT; Cresti 2000, Moneglia & Raso 2014), developed by Emanuela Cresti throughout decades of empirical studies conducted on the LABLITA spoken corpora. L-AcT is an extension of Austin's (1962) Speech Act Theory, and it is a pragmatic theory of speech that acknowledges the paramount role that prosody plays in spoken language. Accordingly, the unit of reference for spoken language is the minimum stretch, within the speech continuum, provided with pragmatic and *prosodic* autonomy, i.e. interpretable as a speech act - the utterance. Prosody is responsible for conveying the illocutionary values that the locutive content of the utterance can assume in speech. In fact, the same (sequence of) word(s), if pronounced with different intonations, can fulfil different illocutions.

Moreover, as it was mentioned in 1.1, prosodic breaks segment the speech continuum into utterances and intonation units within them: terminal, i.e. perceived as conclusive, prosodic breaks signal boundaries between utterances (they are marked by '//' according to the C-ORAL transcription norms), while non-terminal, i.e. perceived as continuing, breaks signal boundaries between intonation units that are part of the same utterance. See the following example, extracted from the Informal C-ORAL-BRASIL corpus:

(1) bfamdl04 [99-108][1]audio (two housekeepers are chatting while working; one of them misidentifies Murano glasses as «Urano glasses»)

*KAT: *o quê* //
*SIL: *copos* // *copos de Urano* / *que tem aí* //
*KAT: *copos de quê* //
*SIL: ***Urano*** //
*KAT: ***Urano*** //
*SIL: *é* // ***Urano*** // ***Urano*** // *é um negócio que tem* / *que es fazem na Itália* / *que custa caríssimo* //

*KAT: what //
*SIL: glasses // Urano glasses / that are right there //
*KAT: glasses of what //
*SIL: Urano //
*KAT: Urano //
*SIL: yeah // Urano // Urano // it's a stuff there / that they make in Italy / that's very expensive //

Example (1) is made up by an exchange of ten utterances (as it is possible to count ten terminal breaks): eight of them are formed by one intonation unit only, while the other two are formed by two and three different intonation units, respectively, as the non-terminal breaks within them signal. There are four utterances (in bold) that display the same locutive content *Urano*, but pronounced with different intonations, conveying different illocutions: a confirmation, an expression of disbelief, and two conclusions (performed with different attitudes), respectively (see Raso 2012 for a more detailed description of the four prosodic profiles).

Starting from these theoretical assumptions, then, it is not surprising that the C-ORAL corpora were and are designed with a set of specifications that make them truly adequate tools for the analysis of spoken language: beside the transcripts

1. The form of citation for any utterance or file from any C-ORAL corpus follows the same criteria: the first letter stands for the language (*b* for BP, *i* for Italian, *e* for Spanish, *f* for French, *p* for European Portuguese); *fam* stands for *private/familiar* context (vs. *pub* for public); *dl*, *cv*, and *mn* stand for the interaction typology: dialogue, conversation, and monologue, respectively; then the number of the file within the corpus follows, and finally that of the reported utterance(s). Starred abbreviations correspond to the speakers' acronyms.

(in CHAT format, MacWhinney 2000) that include the notation of prosodic breaks, all corpora provide the audio files (in .wav format), the text-to-speech alignment (through the WinPitch software: Martin 2015), the metadata with the sociolinguistic information of each recording session, and the PoS tagging (for the BP corpora, trough one version of the parser PALAVRAS, specifically adapted to BP: Bick 2012). Beside sharing the same specifications, the C-ORAL resources follow the same architecture: they include four different corpora (Informal, Formal, Telephonic, Media; see 2.1), organized into the same subsections/thematic domains. In this way, the comparability between them is ensured, enabling crosslinguistic studies —one of the main purposes of the C-ORAL projects within and outside Europe.

Specific characteristics of the C-ORAL-BRASIL resources compared to other C-ORAL corpora are detailed in the following sections.

2.1 C-ORAL-BRASIL corpora specifications and description

In Table 1, C-ORAL-ROM and C-ORAL-BRASIL sizes are compared, according to different measures (number of recordings, duration, number of words, of units of reference, of speakers). The C-ORAL-ROM corpora have an approximate size of 300,000 words each, while there is some variation with regards to the number of speakers and number of recordings, depending on each specific team (for details see: Moneglia 2005). The highest degree of comparability is found between the Italian and the BP corpora, due to a closer scientific relationship between the two teams.

Table 1. C-ORAL-ROM and C-ORAL-BRASIL sizes compared

Language	.wav files	duration (hours)	# words	# utterances	# speakers
French	206	26	295,803	21,010	305
Italian	204	36	310,969	35,446	451
European Portuguese	152	29	317,916	34,067	261
Spanish	210	31	333,482	30,256	410
TOTAL C-ORAL-ROM	772	122	1,258,170	120,779	1,427
Brazilian Portuguese	393	45	498,051	62,505	739

The C-ORAL-BRASIL corpora reached almost 500,000 words, i.e. a 66% size increase compared to the rest of the C-ORAL corpora. Special attention was paid to achieving the greatest possible diaphasic variation in the C-ORAL-BRASIL resources, aiming at obtaining full representativeness and balance.

The Informal corpora are organized into three interactional typologies: 2/3 of dialogues and conversations (two or more than two active participants, respectively), and 1/3 monologues (one active participant only). The recorded sessions are also divided into private/familiar (3/4) and public (1/4) contexts. In the Spanish, French and European Portuguese corpora, these two contexts were discriminated by simply evaluating whether the interaction occurred in a private or in a public place, while in the Italian corpus a more fine-grained evaluation was made, not only of the place, but also of the kind of relationships between the participants. The C-ORAL-BRASIL team decided to use, as a discriminating factor for the context assessment, the role of the speakers within the communicative situation, plus a careful analysis of the topics and/or activities performed by them (for an extensive explanation see: Mello 2014).

The 139 recorded sessions of the Informal corpus include a very diverse range of situations, such as a personal training session, an amateur football game, a visit to a patient in a hospital, the make-up session of a drag-queen before her show, a mother-to-daughter cooking lesson, among many others. [2]

The Informal BP corpus has been published in 2012 (Raso & Mello 2012), and it is available for download, as all the C-ORAL-BRASIL concluded resources, at <www.coral-brasil.org> (> Corpora).

The difference between formal and informal interactions is quite intuitive but not so simple to be precisely described. While it is difficult to set a closed repertory of informal situations, as they are quite unpredictable, sociolinguistic studies set a list of situations that can be considered formal (Gadet 2000). Therefore, the Formal C-ORAL corpora cover professional interactions, conferences, religious functions, and interactions occurring in the academic or law domains of usage (for further details on the general architecture of C-ORAL corpora see Moneglia & Martin 2005: 39), all recorded in natural context.

The Formal in Natural Context BP corpus comprises 74 different recordings divided into 8 main domains (in italics in the list below, where some examples are provided as well):

a. *Business*: e.g. a man receiving consultancy from his bank manager about how to make an investment; a management meeting in a hospital;

2. Details in Raso & Mello (2012).

b. *Conference*: from a variety of scientific areas or topics, e.g. inorganic chemistry, mathematics, health care, water governance, among others;
c. *Law*: e.g. reports at police station; law hearings and trials (especially at Labour court); legal consultancies with lawyers;
d. *Political Debate* and *Political Speech*: e.g. a mayor speech at a public gym inauguration; a trade union meeting; debates in city councils; public speeches and debates during the 2014 election campaign;
e. *Preaching*: masses, evangelical cults, Kardecist (spiritist) and other meetings, according to the actual proportion of the main religious confessions reported by the official *IBGE* census;[3]
f. *Professional Explanation*: e.g. a training session of recently hired workers in a food industry; a pedagogic meeting with the students' parents in a school; a job interview; the presentation of professional courses offered by a private institution; the explanation of a project to a client by an interior designer, among others;
g. *Teaching*: e.g. geography, semantics, aeronautics, sociology of religion, vocal technique, linguistics, odontology classes.

As it can be seen by the list above, a high degree of diaphasic variation was achieved within the Formal corpus as well.

In the Media corpora, the recordings cover a diverse range of thematic domains, both from television and radio shows, in balanced proportions (see Moneglia 2005 for details on the balancing criteria): Interview, Meteorology, News, Documentary, Scientific Press, Sport and Talk Show. The C-ORAL-BRASIL Media corpus has a 132% enlarged size compared to the C-ORAL-ROM ones: the internal balancing was preserved, but a new Extra section has been added, in which new formats (cooking and interior design shows, for example) or exceeding recordings are collected.

The Telephone corpora originally included human and human-machine interactions. Due to the current obsolescence of the latter, C-ORAL-BRASIL Telephone corpus doesn't cover such interactions, and rather distinguishes private and public telephonic exchanges, according to whether they take place between relatives and/or friends, approaching casual and familiar topics, or between people in workplaces, approaching less informal topics. Nonetheless, the Telephonic corpus represent an informal variety of spoken BP.

3. IBGE is the Brazilian Institute for Geography and Statistics. According to its 2010 population census, 65% of the population declares to be Catholic, 22,3% Evangelical, 8% not religious, 2% kardecist, and 2,7 % of other confessions.

2.2 informal, formal, telephonic, and media corpora details

An extensive bibliography already exists on the C-ORAL-BRASIL Informal corpus (Raso & Mello 2012).[4] Here, we only recall that the corpus has a size of 208,130 words for a total of 139 recorded sessions, and that it is representative of the Minas Gerais State diatopy, mainly from its capital city Belo Horizonte.

The same diatopic variety is the main one in the C-ORAL-BRASIL Formal in Natural context and Telephonic corpora (to be published soon). The Formal corpus has a size of 121,396 words, for 74 recorded sessions divided into sections corresponding to the usage domains mentioned in 2.1, while the Telephonic one is a much smaller resource, with a total of 31,308 words. Tables 2 and 3 detail the sizes of each section of both corpora:

Table 2. C-ORAL-BRASIL Formal in Natural Context size per section

Natural context	.wav files	# words	# utterances
Business	4	10,635	1,245
Conference	9	14,195	1,025
Law	9	15,994	2,267
Political debate	12	15,571	1,285
Political speech	15	18,050	1,398
Preaching	9	12,852	1,036
Professional explanation	9	16,219	1,734
Teaching	8	16,291	1,259
Total	74	119,807	11,249

Table 3. C-ORAL-BRASIL Telephonic corpus sizes per section

Telephonic	.wav files	# words	# utterances
Private	50	25,553	4,559
Public	29	5,755	1,291
Total	79	31,308	5,850

4. For an updated list of the publications see: <www.c-oral-brasil.org->Publications>.

Finally, the Media corpus has a size of 139,647 words —Table 4 shows how they are distributed across its different formats. In this resource, other diatopic varieties of spoken BP appear.

Table 4. C-ORAL-BRASIL Media corpus sizes per section

Media	.wav files	# words	# utterances
Interview	9	15,506	1,492
Metereology	1	232	11
News	9	6,096	399
Documentary	29	23,530	2,542
Scientific Press	12	13,233	1,062
Sport	7	12,234	1,075
Talk Show	18	44,088	3,838
Extra	16	24,728	2,586
Total	101	139,647	13,005

2.3 methodological practices

The C-ORAL-BRASIL project adopts a set of validating procedures to ensure the reliability of the compiled resources at different levels.

For what concerns the prosodic breaks annotation, statistical validation of the internal agreement between annotators is run through the Kappa test (Fleiss 1971). The scores of the test vary between 0 (chance agreement) and 1 (total agreement), and within the C-ORAL-BRASIL (for C-ORAL-BRASIL I specifically, see Mello *et al.* 2012) the acceptability threshold is set at 0.8 for terminal breaks (starting from 0.8, scores correspond to quasi-total agreement), and at 0.6 for non-terminal breaks (scores between 0.6 and 0.7 correspond to substantial agreement).

A noteworthy methodological practice is that of running the Kappa test not only at the final stage of transcription, but throughout different stages of any resource's compilation. As it was mentioned in 1.1, the complexity of the compilation of a spoken corpus requires much effort and an articulated teamwork. The C-ORAL-BRASIL research group is composed by a large group of undergraduate students, that work under the supervision of senior, more experienced researchers —graduate students and professors, mostly, but also undergraduate students with a longer, pre-

vious research experience within the group. The exchange of undergraduate students working within the C-ORAL-BRASIL is constant and sometimes very fast, so that the research group has its composition renewed. In order to develop any research within the LEEL lab, any researcher (junior and senior) receives a four to six months training in speech segmentation and transcription. Through the supervision of a senior prosodic annotator, the perception of the distinction between terminal and non-terminal breaks by the new researchers is trained on actual spontaneous recordings, encompassing all interactional types. The trainer is also responsible for checking that all the C-ORAL-BRASIL transcription criteria are followed by the new transcribers. At the end of the training, a first Kappa test is run, to check which transcribers can actually start working, and which ones need to continue the training process. It is of paramount importance that a good agreement exists between transcribers to avoid random annotation of prosodic breaks.

Aiming at an easy readability of its transcripts, the C-ORAL-BRASIL corpora are transcribed following mostly an orthographic criterion. Nonetheless, a series of non-orthographic criteria was set, in order to document a few morpho-phonetic and grammaticalization phenomena that could be of interest for the study of spoken BP. These include apheretic forms (*brigado* for *obrigado* «thanks», *cabou* for *acabou* «it's the end», and so on), cliticization of subject pronouns (*es* for *eles* «they», *cê* for *você* «you», among others), loss in verbal morphology or other grammaticalization processes (grammaticalized apheretic forms of the verb *estar*, among others; see Mello *et al.* 2012 for an extensive description of all non-orthographic criteria in C-ORAL-BRASIL I).

Several compilation phases follow the transcription one: every transcript passes through two revision processes, before and after the alignment to the sound signal, performed through WinPitch. For the C-ORAL-BRASIL II resources, two more revision processes were added, so that the transcripts are checked four times, for prosodic annotation and transcription criteria, twice before and twice after being aligned —of these, the latest is specifically focussed on the prosodic segmentation. It is important to notice that expert and more skilled researchers are responsible for the revision phases, and that the team of transcribers and revisers are composed by different people, minimizing possible biases related to specific researchers.

A second Kappa test for the prosodic breaks' segmentation is run closer to the final stages of the compilation of the resource. The final score in the C-ORAL-BRASIL I corpus was 0.87 (details in Mello *et al.* 2012), while for C-ORAL-BRASIL II corpora, the scores were 0.83 for the Telephonic corpus, 0.84 for the Formal, and 0.86 for the Media corpus —an excellent inter-transcriber agreement was achieved.

Throughout the compilation of the C-ORAL-BRASIL II resources, another validation test was run with respect to transcription criteria. The main purposes of the test were to verify the reliability of the transcripts for what concerned:

a. total correspondence between actual pronunciation and transcribed words;
b. compliance to orthographic criteria;
c. compliance to non-orthographic criteria.

For the corpus C-ORAL-BRASIL I, these validation procedures were limited to a 5% sample on the total amount of utterances, setting a threshold error rate of 5%.

For the C-ORAL-BRASIL II resources, two teams of two trained, expert transcribers under the supervision of a senior researcher randomly sampled approximately 9% on the total of utterances of the corpora. A total of 31 criteria were verified in the sample: for more frequent phenomena, random samples were enough to ascertain the compliance to the respective criteria, while less frequent phenomena required targeted queries to be checked (details on the validation methodology are found in Santos & Raso in preparation). The same threshold for error rate of 5% was set for the validation within the C-ORAL-BRASIL II resources, and only after this step the team passed to the final revision before publication.

One of the main characteristics of the C-ORAL-BRASIL resources is, as shown before, its concern with the widest possible diaphasic variation. For the Informal and Formal in Natural Context corpora, the achievement of such a goal was possible thanks to non-invasive, high-quality equipment used by the LEEL members in the recording process: wireless lapel microphones (Sennheiser Evolution EW100 G3 wireless kits) and a mixer (Behringer XEXYX 1222 FX) allowed the speakers to have freedom of movements and more easily maintain the naturalness of the interactional situations. Besides, the high-resolution digital recorders used in the project (MARANTZ PMD660 Professional Solid-State Recorder or TASCAM DR-100) ensured the viability of precise acoustic analysis of a large part of the corpus.

Another methodological improvement found in the C-ORAL-BRASIL II resources in comparison to the C-ORAL-BRASIL I regards the measurement of the acoustic quality throughout the corpora. A fine-grained scale of acoustic quality is used to label each recording session of the C-ORAL-BRASIL resources: A - AB - B - BC - C - D are the six grades of the scale, from the highest to the lowest acoustic quality tag. For the C-ORAL-BRASIL I corpus, the acoustic quality tagging was made by specific researchers, through hearing impressions and inspection of the readability of the spectrograph, considering the following criteria: (a) microphone response;

(b) acoustic analysability; (c) rate of speech overlapping; (d) rate of background noise; (e) f0 computability and f0 tracking reliability; (f) audibility (cf. Raso 2012a: 74).

For the C-ORAL-BRASIL II corpora, a new, semi-automatic acoustic quality evaluation protocol was implemented. Researchers could count on a set of *ad hoc* Praat (Boersma & Weenink 2019) scripts, checking samples for speech overlapping, noise-signal ratio, f0, f1 and f2 formants analysability, background noise rate, among others (see Vieira *et al.* in preparation for details). All these parameters received different weights, for the script to calculate the final acoustic quality tag. Only 15 recording received the C tag within the C-ORAL-BRASIL II resources, while the most part of them displays A to BC tags.

3. OTHER C-ORAL-BRASIL RESOURCES: THE MINICORPORA

The C-ORAL-BRASIL project has compiled and is compiling a set of resources aimed at studies specifically focussed on information structure and its interfaces.

According to the L-AcT, intonation units within the utterances can convey information units, corresponding to different pragmatic/communicative functions —conveyed, one more time, by specific *prosodic* profiles, together with specific distributional properties within the utterance. Information units are divided into two main categories: *textual* units, that make up the semantic and syntactic content of the utterance, and *dialogic* units, that are directed to the interlocutor and correspond to what is called discourse markers within different approaches. In Table 5 we present a sketch of the main information units (and their tags) according to the L-AcT:

Table 5. Information units according to L-AcT (adapted from Moneglia & Raso 2014, p. 490-491)

Type	Name	Tag	Function
Textual	Comment	COM	It carries the illocutionary force of the utterance, being the only necessary and sufficient information unit.
	Topic	TOP	It establishes a domain of application for the illocution of the Comment.
	Appendix of Comment	APC	It integrates the text of the Comment, adding information to it.
	Appendix of Topic	APT	It integrates the text of the Topic, adding to it a delayed information.
	Locutive Introducer	INT	It signals a following meta-illocution, such as reported speech, emblematic exemplification, or spoken thought.

	Parenthesis	PAR	It inserts information with a metalinguistic function, providing instructions on how to interpret the utterance or part of it.
	Multiple Comments	CMM	They constitute a chain of Comments forming an illocutionary pattern combining different illocutions for the performance of one conventionalized rhetoric effect.
	Bound Comments	COB	They form a sequence of Comments produced by a progressive juxtaposition that follows the flow of thought.
Dialogic	Incipit	INP	It opens the communicative channel, starting a dialogic turn or an utterance.
	Conative	CNT	It pushes the listener to take part of interaction.
	Phatic	PHA	It controls the status of the communicative channel, ensuring its maintenance.
	Allocutive	ALL	It specifies to whom the message is directed; it has an empathic function.
	Expressive	EXP	It works as an emotional support, stressing the sharing of a social affiliation.
	Discourse Connector	DCT	It connects different parts of the discourse, indicating its continuation.

In order to enable crosslinguistic studies on information structure, a group of comparable, informationally annotated resources was created, through the selection of representative samples of the Informal C-ORAL corpora. These resources are known as the C-ORAL minicorpora, and the first two of them were the Italian and the BP ones, available for online query at the DB-IPIC *Database for Information Pattern Interlinguistic Comparison* platform (<http://www.lablita.it/app/dbipic>), developed by the LABLITA lab (Panunzi & Mittmann 2014).

The two minicorpora comprise 20 recording sessions and reproduce the same architecture of their reference corpora in what concerns the proportion between monologic and dialogic interactions, and the balancing between private and public contexts. They all maintain the specifications of the C-ORAL corpora (transcription with notation of prosodic breaks, audio files, text-to-speech alignment, etc.), but are provided with informational tagging, which was made manually by a team of trained annotators. In (2), we repeat example (1) with the informational tagging (bfamdl04 is part of the BP minicorpus):

(2) bfamdl04 [99-108]

*KAT: *o quê* //=COM=
*SIL: *copos* //=COM= *copos de Urano* /=COM= *que tem aí* //=APC=
*KAT: *copos de quê* //=COM=

*SIL: ***Urano*** //=COM=
*KAT: ***Urano*** //=COM=
*SIL: *é* //=COM= ***Urano*** //=COM= ***Urano*** //=COM= *é um negócio que tem* /=SCA= *que es fazem na Itália* /=TOP= *que custa caríssimo* //=COM=

At present, a new version of the BP minicorpus has been published, which passed through a thorough revision of the informational tagging by a new team of trained annotators. The new BP minicorpus is fully available for download at <www.c-oral-brasil.org>, i.e. their transcripts (.txt), audio files (.wav), alignment files (.xml; accessible through WinPitch, which itself can be freely downloaded at <www.winpitch.com>), and metadata.

These resources are a precious tool for studies focussed on information structure and its interfaces and, seeking to widen the scope of crosslinguistic comparison, new informationally tagged minicorpora were compiled or are being compiled within the C-ORAL-BRASIL project.

3.1 other minicorpora

An American English minicorpus has been extracted from the Santa Barbara Corpus of Spoken American English (Du Bois *et al.* 2000-2005) and adapted to the transcription criteria of the C-ORAL family (Cavalcante & Ramos 2016). The sampling criteria included good acoustic quality, and the search for the same architecture of the other minicorpora (20 recording sessions, 2/3 dialogic vs. 1/3 monologic interactions, 3/4 private vs. 1/4 public contexts). This resource is particularly relevant because it widens the scope of crosslinguistic comparison outside the Romance domain.

Returning to BP, a Telephonic minicorpus has been recently compiled as well. It comprises 27 recorded sessions form the public and the familiar contexts. This resource has been designed to be comparable to the dialogic face-to-face interactions of the BP minicorpus, aiming at the contrastive study of the two different diamesic varieties.

Both the American English and the Telephonic minicorpora are already available for download at the C-ORAL-BRASIL website.

Other minicorpora are being compiled by now:

1. a new Informal Italian minicorpus, through a new selection of recordings from the Italian C-ORAL-ROM that aim to improve the overall acoustic quality of the resource; this minicorpus is passing through a revision process and will be soon available for download;

2. an Angolan Portuguese minicorpus, from a collection of recordings realized in July 2018 by a team from the LEEL lab in collaboration with the *Projeto Libolo* from the São Paulo State University (Rocha *et al.* 2019); so far, only the selection of the 20 recordings that fit the C-ORAL architecture and acoustic quality requirements was accomplished;
3. a minicorpus extracted by the C-ORAL-ESQ corpus, a corpus of BP spoken by patients affected by schizophrenia, which is itself being compiled by now; this resource aims at exploring the linguistic effects (specifically, on information structure) of the disease (see Rocha 2019 for details).

4. FINAL REMARKS

Throughout almost a decade of hard teamwork, the C-ORAL-BRASIL project has compiled and is compiling a variety of truly adequate resources for the study of spoken BP, and other varieties of Portuguese or languages —enlarging, thus, the scope of crosslinguistic comparison in all levels of linguistic analysis. Beside complying with the technical demands that the compilation of spoken corpora requires, through a set of specifications and the continuing improvement of related methodological practices, the C-ORAL-BRASIL resources are made fully available (with all their specifications) to the scientific community as soon as the compilation process is completed. For many reasons, then, its resources can be included within the golden standard models for the compilation of spoken corpora within and outside Brazil.

Lúcia de Almeida Ferrari
Universidade Federal de Minas Gerais
ferrari.lu@gmail.com
ORCID 0000-0002-9855-0646

Giulia Bossaglia
Universidade Federal de Minas Gerais
giulia.bossaglia@gmail.com
ORCID 0000-0001-8839-3088

BIBLIOGRAPHIC REFERENCES

Austin, L. J. (1962) *How to Do Things with Words*, Oxford, Oxford University Press.

Berber Sardinha, T. (2004) *Lingüística de corpus*, Barueri, Manole.

Bick E. (2012) «A anotação gramatical do C-ORAL-BRASIL», in T. Raso & H. Mello (eds.), *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal*, Belo Horizonte, Editora UFMG, p. 223-254.

Boersma, P. & D. Weenink (2019) *Praat: doing phonetics by computer* [Computer program], Version 6.1.03. [Online: <http://www.praat.org>, accessed: 2019-09-01.]

Cavalcante, F. A. & A. C. Ramos (2016) «The American English spontaneous speech minicorpus: architecture and comparability», *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, v. 3.2, p. 99-124.

Cresti, E. (2000) *Corpus di Italiano Parlato*, vol. I-II, CD-ROM, Firenze, Accademia della Crusca.

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

Davies, M. & M. Ferreira (2006) *Corpus do português*: 45 milhões de palavras, 1300s--1900s. [Online: <http://www.corpusdoportugues.org>, accessed: 2019-09-01.]

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

Fleiss, J. L. (1971) «Measuring nominal scale agreement among many raters», *Psychological Bulletin*, v. 76, p. 378-382.

Gadet, F. (2000) « Vers une sociolinguistique des locuteurs », *Sociolinguistica*, 14, p. 99-103.

Jefferson, G. (1984) «On the organization of laughter in talk about troubles», in J. M. Atkinson & J. Heritage (eds.), *Structures of social action. Studies in conversation analysis*, Cambridge, Cambridge University Press, p. 346-369.

MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk* 3rd Edition, Mahwah / NJ, Lawrence Erlbaum Associates.

Martin, P. (2015) *Winpitch Pro* W8. v 7.2.00, Pitch Instruments. [Online: <http://www.winpitch.com>, accessed: 2019-09-01.]

Mello, H. R. (2014) «Methodological issues for spontaneous speech corpora compilation: The case of C-ORAL-BRASIL», in T. Raso & H. R. Mello. (eds), *Spoken Corpora and Linguistic Studies*, Amsterdam/Philadelphia, John Benjamins, p. 27-68.

Mello, H., T. Raso, M. M. Mittmann, H. P. Vale & P. Côrtes (2012) «Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação», in T. Raso & H. R. Mello (eds.), *C-ORAL - Brasil I: Corpus de referência do português brasileiro falado informal*, Belo Horizonte, Editora UFMG, p. 125-176.

Moneglia, M. & P. Martin (2005) «The C-ORAL-ROM resource», in E. Cresti, & M. Moneglia (eds.), *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*, Amsterdam/Philadelphia, John Benjamins, p. 1-70.

Moneglia, M. & T. Raso (2014) «Notes on Language into Act Theory», in T. Raso & H. R. Mello (eds), *Spoken Corpora and Linguistic Studies*, Amsterdam/Philadelphia, John Benjamins, p. 468-495.

Raso T. (2012a) «O corpus C-ORAL-BRASIL», in T. Raso & H. R. Mello (eds.), *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*, Belo Horizonte, Editora UFMG, p. 55-90.

— (2012b) «O C-ORAL-BRASIL e a Teoria da Língua em Ato», in in T. Raso & H. R. Mello (eds), *C-ORAL-BRASIL I: Corpus de referência do Português Brasileiro falado informal*, Belo Horizonte, Editora UFMG, p. 91-123.

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

Raso T., H. Mello & L. A. Ferrari, *C-ORAL-BRASIL II*, forthcoming.

Rocha, B. N. R. (2019) «O corpus C-ORAL-ESQ e a estrutura informacional da fala de pacientes com esquizofrenia», *Working Papers em Linguística*, 20(1), p. 212-238.

Rocha, B. N. R., H. Mello & T. Raso (2019) «Para a compilação do C-ORAL--ANGOLA: um corpus de fala espontânea informal do português angolano», *Filologia e Linguística Portuguesa* (Online), v. 20, p. 139-157.

Santos, S., T. Raso, T. Arantes, A. Neves, A. Silva & T. Viana, «Manual validation of transcription criteria of the C-ORAL-BRASIL II language resource: assessed criteria, methodology, and results», forthcoming.

Schmidt, T. (2004) «Transcribing and annotating spoken language with EXMARaLDA», in *Proceedings of the LREC-Workshop on XML based richly annotated corpora*, Lisbon/Paris, ELRA.

Vieira, M. A., E. Ramos & T. Raso, «A practical protocol for audio quality evaluation in spontaneous speech corpora», forthcoming.