Journal of Speech Sciences 7(2): 65-77. 2018.
Available at: http://revistas.iel.unicamp.br/joss

THE C-ORAL-BRASIL PROJECT: VARIED RESOURCES FOR THE STUDY OF SPOKEN BRAZILIAN PORTUGUESE

BOSSAGLIA, Giulia[1]*
FERRARI, Lúcia de Almeida[2]*

[1]Universidade Federal de Minas Gerais
[2]Universidade Federal de Minas Gerais

Abstract: *In this paper, we present different resources for the study of spoken Brazilian Portuguese, developed within the C-ORAL-BRASIL project. The C-ORAL-BRASIL stemmed from the European C-ORAL-ROM project (Cresti and Moneglia, 2005), which has compiled spoken corpora of Italian, French, Spanish, and European Portuguese. The corpora of the C-ORAL family represent adequate tools for the analysis of spoken language, for they are provided not only with the transcripts of the recorded sessions (with prosodic breaks' annotation), but also with their audio files and the text-to-speech alignment (with the WinPitch software: Martin, 2003). So far, the C-ORAL-BRASIL project has published the C-ORAL-BRASIL I (Informal corpus: Raso and Mello, 2012), while the C-ORAL-BRASIL II (to be published) comprises a Formal corpus (Natural context), a Media corpus, and a Telephonic corpus. Besides these resources, a set of informationally tagged comparable minicorpora (representative samples of the aforementioned corpora) are already available or in preparation, enabling (cross-linguistic) studies focused on information structure and its interfaces.*

Keywords: *spoken corpora; spontaneous speech; Brazilian Portuguese; C-ORAL-BRASIL.*

1 The C-ORAL family of spoken corpora: a brief review

The C-ORAL-ROM project (Cresti and Moneglia, 2005) aimed at compiling comparable resources for the study of spontaneous speech of some European Romance languages, resulting in the publication of four spoken corpora: the Italian corpus by the LABLITA lab at Florence University (https://www.letterefilosofia.unifi.it/vp-309-lablita.html), the French corpus by the DELIC project at the Université de Provence (http://www.elda.org/en/proj/coral/corpus_delic.html), the Portuguese corpus by the CLUL lab at Lisbon University (http://www.clul.ulisboa.pt/en/resources-en), and the Spanish corpus by the Linguistics lab at the Universidad Autónoma de Madrid (Moneglia and Martin, 2005).

All corpora were planned as effective tools for the study of spoken language: besides the transcripts of the recording sessions, they provide the audio files and the text-to-speech alignment through the WinPitch software (Martin, 2003; see Figure 1).

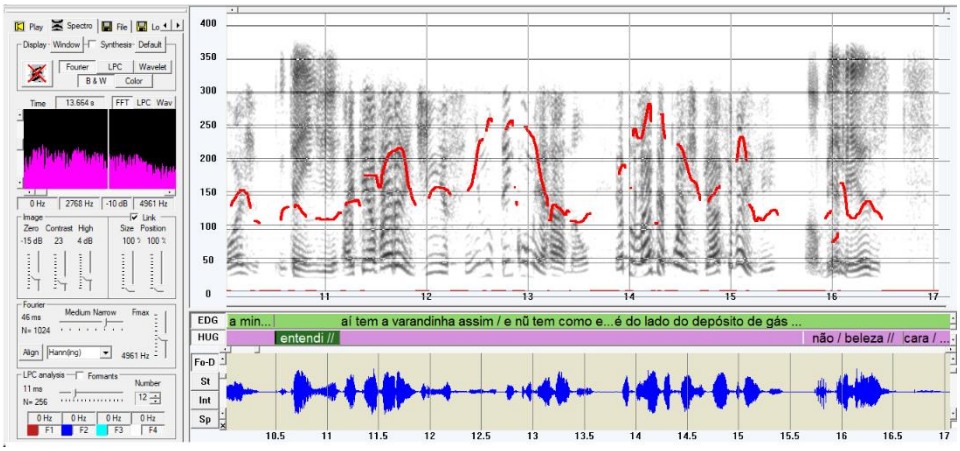

Figure 1: WinPitch window showing text-to-speech alignment

*Both corresponding authors: giulia.bossaglia@gmail.com; ferrari.lu@gmail.com

Through the text-to-speech alignment, as it can be seen in Figure 1, it is always possible to access the transcript (in CHAT format: MacWhinney, 2000), the corresponding audio, and the spectrogram. The prosodic information is of paramount importance for the study of spoken language, and prosody is held as a primary means to convey meaning (on many levels: semantic, pragmatic, and so on) in speech.

All C-ORAL corpora follow the Language into Act Theory (L-AcT: Cresti, 2000; Moneglia and Raso, 2014), a corpus-driven framework that assumes a prosodic and pragmatic unit of reference for spoken language: the utterance. The utterance corresponds to a speech act (Austin, 1962), i.e. to a locutionary act endowed with illocutionary force – this is conveyed by a specific prosodic nucleus, without which the utterance could not be interpreted as pragmatically autonomous. Boundaries between utterances are prosodically marked too, by terminal, i.e. perceived as conclusive, prosodic breaks; non-terminal breaks, perceived as continuing, signal the boundaries between intonation units within the utterance (according to L-AcT, each segmentable intonation unit conveys a specific information unit, see 2.3.1). Accordingly, all C-ORAL transcripts are provided with the annotation of prosodic breaks, which is made manually by teams of trained transcribers (see Barbosa and Raso, 2018, and Teixeira, Barbosa and Raso, 2018 on a semi-automatic tool for the detection of prosodic breaks, developed on the basis of the C-ORAL-BRASIL corpora).

The C-ORAL-BRASIL project stemmed as the non-European branch of the C-ORAL family at the LEEL lab (http://www.letras.ufmg.br/leel/) at the University of Minas Gerais (Brazil). Due to a closer collaboration between the LABLITA and LEEL labs, the C-ORAL-BRASIL corpora are closer in architecture and comparability to the Italian ones.

The global architecture of the C-ORAL resources comprises four corpora: the Informal and Formal corpora, Media, and Telephone (see Table 1, adapted from Moneglia and Martin, 2005, p. 39). The recordings of the Informal corpus have an average size of approximately 1,500 words (short texts) and 4,500 words (long texts); those of the Formal and the Media corpora, approximately 3,000 words; the Telephone corpus has an established upper limit of 1,500 words, but no lower limit. The texts of the Informal, Formal, and Media corpora are organized according to three different interactional types (monologue, dialogue, conversation).

Table 1: C-ORAL-ROM corpora structure with codes for text type.

	Field	**Subfield**	
Informal	fam (family-private)	mn	monologue
	pub (public)	dl	dialogue
		cv	conversation
Formal	nat (natural context)	ps	political speech
		pd	political debate
		pr	preaching
		te	teaching
		pe	professional explanation
		bu	business
		co	conference
		la	law
Media	med (media)	nw	news
		mt	weather forecast
		in	interview

		rp	documentary
		sc	scientific press
		sp	sport
		ts	talk show
Telephone	tel (telephone)	pv	private
		mm	man machine

Metadata of each recording session are provided as well, including information about context of the recorded session (private-familiar vs. public, plus other relevant information on the communicative situation), size (word number and length), speakers' sociolinguistic profiles (age, sex, education, profession, birthplace), acoustic quality, among others.

Following the same theoretical and methodological framework, the more recent C-ORAL-BRASIL corpora could benefit from technological and methodological advances, as it will be shown throughout the following sections.

2 The C-ORAL-BRASIL Project

The C-ORAL-BRASIL resources comprehend the C-ORAL-BRASIL I (Informal corpus) and the C-ORAL-BRASIL II (Formal in Natural Context, Media and Telephone corpora). These corpora stand out in size within the C-ORAL family, not only for the overall word number and number of segmented utterances, but also for each part of the general architecture, as it is shown in Tables 2 and 3:

Table 2: C-ORAL-ROM vs. C-ORAL-BRASIL overall sizes

	.wav files	**Duration (hours)**	**Words**	**Speakers**
French	206	26	295,803	305
Italian	204	36	310,969	451
Portuguese	152	29	317,916	261
Spanish	210	31	333,482	410
Brazilian Portuguese	393	45	500,481	741

Table 3: C-ORAL-ROM vs. C-ORAL-BRASIL sizes (utterance number) per section[1]

	French	**Italian**	**Portuguese**	**Spanish**	**Brazilian Portuguese**
Informal	10,517	23,805	21,949	21,618	31,442
Formal in Natural Context	2,880	5,503	5,542	4,138	10,599

[1] Corpora comparability depends more on the number of reference units (i.e. utterances) than of words. Since the utterance corresponds to a speech act, highly actional communicative situations (dialogues and conversations) can display a high number of utterances but a low word number, while less actional interactions (monologues) tend to have less utterances, but higher word numbers. Thus, word number is not the best parameter to measure comparability between corpora, and the comparison between each section of the C-ORAL-ROM and C-ORAL-BRASIL resources is made following the utterance number as a measure.

Media	3,633	6,138	1,570	4,500	13,005
Telephone	3,980	4,642	4,788	5,332	5,850
Total	21,010	35,446	34,067	30,256	60,896

The increase in size does not challenge the comparability within the family, since the balance of the C-ORAL architecture is maintained.

Thanks also to later technological advances, the C-ORAL-BRASIL II resources are characterized by a few methodological improvements compared to the rest of the family. Acoustic quality was ensured by high technology equipment (PMD660 Marantz digital recorders and Sennheiser Evolution EW100 G2 wireless kits), while the non-invasive clip-on microphones helped the maintenance of the spontaneity of the situations, crucial to achieve the high degree of diaphasic variation, which was the primary goal in the compilation of the Informal and Formal corpora (see Mello, 2014 for an extensive discussion on the importance of diaphasic variation in designing the architecture of spoken corpora).

An even greater attention to the acoustic quality of the recordings was paid in C-ORAL-BRASIL II corpora with respect to C-ORAL-BRASIL I. Recordings were labelled according to a fine grained, six-degree scale: from A (very high quality) to D (low quality), with intermediate AB, BC degrees. The acoustic quality labelling process ceased to be done according only to the researchers' judgement based on a few parameters (among others: amount of speech overlapping, f0 computability by the software, noise rate, and phonetic analyzability). For C-ORAL-BRASIL II corpora, different Praat (Boersma and Weenink, 2016) scripts (Vieira et al., forthcoming) were implemented to scan the .wav files in order to check for the same parameters, but based on objective measurements and in an automatic, much more accurate way.

Additionally, the compilation process included several revision phases (transcription revision, prosodic annotation revision, alignment revision), and for the C-ORAL-BRASIL II resources a new revision phase was added for orthographic and non-orthographic (aphaeretic forms, grammaticalized forms, among others) criteria. Transcription and revision tasks were assigned to specific trained teams, whose internal agreement was measured by applying the Kappa test (Fleiss, 1971) at different stages of the compilation process: before starting and during the transcription process, transcribers' internal agreement in the prosodic annotation (terminal and non-terminal breaks) was statistically validated through the Kappa test, setting a threshold of a minimum 0.6 score for non-terminal breaks and 0.8 for terminal breaks.[2]

One of the final procedures before the publication of the C-ORAL-BRASIL II corpora was the validation of orthographic and non-orthographic transcription criteria: a selected team of three revisers checked samples of each corpus (a random selection of 10 utterances for each text file) for 29 different transcription criteria, setting a maximum of 5% ($p \leq 0.05$) error as acceptable. When some criteria exceeded this error threshold, all texts would be checked again until achieving the targeted score (see Santos and Raso, forthcoming; on validation issues of spoken corpora: van den Heuvel, 2008 and Schiel, 2012).

In the following sections, each of the C-ORAL-BRASIL II resources will be presented in more detail, focusing on relevant methodological issues concerning their compilation.

[2] In the Kappa test, a 0 score means agreement by chance, while a 1 score corresponds to perfect agreement. Details on the results of Kappa test for C-ORAL-BRASIL I are presented in Mello et al. (2012).

Additionally, the informationally tagged minicorpora compiled by the LEEL lab will be introduced as well.

2.1 C-ORAL-BRASIL I: Informal corpus

As the C-ORAL-BRASIL I Informal corpus was published in 2012 (it is ready for download at www.c-oral-brasil.org > Corpora), plenty of descriptions and studies based on it have been available since then (an updated list is provided at www.c-oral-brasil.org > Publications). Therefore, in this section we will only sketch a few of its specifications.

The Informal corpus is composed by 139 recording sessions (a total of 21:08:52 hours; 208,130 words). The main goal of the corpus was to document Brazilian Portuguese (BP) spontaneous (i.e. non-elicited) speech. The main diatopy represented by C-ORAL-BRASIL I is Minas Gerais State's, mostly from its capital city Belo Horizonte, but other varieties are found as well.

Texts are divided into private-familiar (approximately 3/4) vs. public (1/4) contexts, distinguishing between three different interactional types, defined according to the number of active participants in the recorded situation: monologues (one active participant), dialogues (two participants), and conversations (more than two participants). The proportion between dialogic (dialogues + conversations) and monologic interactions is of 2/3 vs. 1/3, respectively.

The Informal corpus is characterized by a significant degree of diaphasic variation, because the greatest diversity of communicative situations was documented: football teammates interacting while playing a game, a drag queen preparing for a show, a realtor driving a client to visit an apartment, a personal training session, among many others. Pursuing the greatest diaphasic variation had as a consequence an almost perfect diastratic balance as well, in what regards sex and schooling of the informants (see Raso, 2012 for details).

Each published recording is made available in all its components: audio file (.wav format), orthographic transcription with prosodic breaks' annotation (.txt and .rtf formats), text-to-speech alignment (.xml format, through WinPitch software), metadata (.txt), and PoS tagged transcripts (through PALAVRAS parser: Bick, 2000).

2.2 C-ORAL-BRASIL II: Formal, Media, Telephone corpora

2.2.1 Formal in Natural Context corpus

The Formal in Natural Context corpus (to be published) is composed by 74 recording sessions (119,807 words); following C-ORAL-ROM architecture, the corpus is divided into different sections that cover various domains of usage (see Table 4).

Table 4: C-ORAL-BRASIL Formal in Natural Context corpus overall sizes.

Natural Context	.wav files	number of words	number of utterances
Business	4	10,851	1,170
Conference	9	17,320	1,151
Law	9	16,107	2,061
Political Debate	12	15,707	1,206
Political Speech	15	16,047	1,135
Preaching	9	12,826	994
Professional Explanation	8	16,247	1,655

Teaching	8	16,291	1,227
Total	74	121,396	10,599

Although the majority of formal speeches tend to be mainly monologic, some dialogues and conversations are also found, due to the high degree of diaphasic variation pursued in the Formal corpus as well: the Business section includes several commercial transactions, the recording of a man receiving advice and information from his bank manager on how to make an investment, among others; in Law, actual court hearings and trials are found (e.g. an AIDS patient requesting pension to social welfare services, a labor law trial about the installation of a camera in the female employees' changing room at a workplace), together with reports filed at a police station; Political Debate and Political Speech sections include, among other situations, a mayor's speech at a public gym inauguration and some public hearings or discourses during campaign rallies.

Globally, the male/female voices' balance mirrors the actual sex differences found in the professional or social groups recorded (60% men and 40% women): in the Business, Conference, Law, Teaching and Professional Explanation sections, the number of words pronounced by male and female voices are well balanced, while in Political Debate, Political Speech and Preaching, the number of words pronounced by men is much higher than women's (72% male vs 28% female in Political Debate; 85% male vs 15% female in Political Speech; 92% male vs 8% female in Preaching).

Another concern of the C-ORAL-BRASIL project was to portray the diversity of Brazilian confessions; therefore, the Preaching section has been balanced according to the to the data of 2010 Census on religious confessions, so that there are 6 catholic masses, 2 evangelical cults and 1 kardecist (spiritist) meeting.[3]

One of the biggest challenges for the compilation of this corpus concerned data collection,
due to the young age of the recording team: they faced difficulties in obtaining consent to the recordings without professors' interposition, since these interactional contexts often involve sensitive information, and public figures such as lawyers, magistrates, doctors, politicians and managers.

2.2.2 Media corpus

The Media corpus (to be published) is composed of 101 recordings (139,647 words). Its internal divisions comply with the C-ORAL-ROM architecture, with the addition of a new section (Extra) that collects the recording surplus, as well as formats that are not found in the C-ORAL-ROM, like cooking and interior design TV shows (see Table 5).

Table 5: C-ORAL-BRASIL Media corpus overall sizes

Natural Context	.wav files	number of words	number of utterances
Interview	9	15,506	1,492
Meteorology	1	232	11
News	9	6,096	399

[3] According to 2010 IBGE's (*Instituto Brasileiro de Geografia e Estatística* "Brazilian Institute for Geographics and Statistics") Census, 65% of the population declared themselves as Roman Catholic, 22,3% Evangelical, 8% not religious, 2% kardecist, and 2,7 % of other confessions.

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

This recording process was actually very simple for the C-ORAL-BRASIL team, as it was only necessary to connect the high-quality recording equipment to the TV/radio, following the schedules of the shows. The biggest issue to be dealt with for this corpus concerns copyright clearance with the broadcasting companies (so far, only one of them has given consent to the use of the data).

The corpus is lightly unbalanced in male/female number of words (61% male and 39% female); actually, Interview and Talk Show present a stronger male voices' prevalence (67% male vs 33% female and 66% male vs 34% female number of words, respectively), and in Sport this discrepancy is even bigger, with 86% of male voices. Those numbers, as a matter of fact, mirror the actual majority of male hosts in this kind of programs.

2.2.3 Telephone corpus

The architecture of the Telephone corpus (to be published) is slightly different from C-ORAL-ROM's. The span of many years between the latter and the beginning of the collection of C-ORAL-BRASIL II resulted in the disappearance of the automatic answering machine (one of the original C-ORAL sections was Human-machine) and in the obsolescence of landlines in favor of mobile phones, even at home. This was one of the biggest challenges for the compilation of this corpus, since the available high quality recording devices were appropriate only for landlines and did not fit mobiles.

The corpus is divided into two sections: Private and Public. The first one includes phone calls between relatives and friends (with topics like health, trips, favors, visits, school projects and so on). In the Public section, the calls were either incoming on a business landline (a beauty salon, an IT company, a bookstore, a physical therapy center and other companies), or directed to business from the researcher's line, in order to record actual public interactions.

The Private section includes 50 recording sessions (25,533 words), while the Public section is composed by 29 recording sessions (5,755 words).

Table 6: C-ORAL-BRASIL Telephone corpus overall sizes

Natural Context	.wav files	number of words	number of utterances
Private	50	25,553	4,559
Public	29	5,755	1,291
Total	79	31,308	5,850

The male/female voices' balance is quite good, with a slight prevalence of women's words in Private (55% female vs 45% male in Private; 49% female vs 51% male in Public).

2.3 The C-ORAL minicorpora for the study of information structure

The compilation of specific resources for the study of information structure goes back to the creation of the DB-IPIC-Database for Information Patterning Interlinguistic Comparison platform by the LABLITA lab (http://www.lablita.it/app/dbipic/; Panunzi and Gregori, 2011), where the Informal section of Italian C-ORAL-ROM corpus is available, together with three comparable informationally tagged minicorpora of Informal spoken Italian, Brazilian Portuguese (thanks to the collaboration with the LEEL lab: Panunzi and Mittmann, 2014), and Spanish (the latest entry in the platform, thanks to the collaboration with the Spanish C-ORAL-ROM team: Nicolás Martinez and Lombán, in press).

These minicorpora are representative samples of their reference corpora of the C-ORAL family, whose architecture they reproduce, in what regards the proportions between private-familiar and public contexts (3/4 vs. 1/4), and dialogic and monologic interactions (2/3 vs. 1/3). Besides all the specifications they share with the C-ORAL corpora (transcripts with prosodic breaks annotation, audio files, text-to-speech alignment, PoS tagging, metadata), these resources were manually tagged for information structure, following the L-AcT theoretical framework.

2.3.1 Information structure: a short premise

According to the L-AcT, each intonation unit segmented within the speech flow conveys a specific information unit (IU). Information units are identified according to their specific prosodic profiles, functions, and distributions within the utterance. The IU's repertory within L-AcT goes beyond the traditional binary opposition between Topic and Comment (or Topic and Focus, Theme-Rheme, and so on), and it is assumed that the informational relations found in spoken language are much more diverse, and prosodically constrained. Following the IPO model (Institute for Perception Research of the University of Eindhoven, t'Hart et al., 2006), L-AcT assumes a correspondence between types of prosodic units and types of IUs, so that the utterance is composed by a prosodic pattern containing at least one root prosodic unit (the Comment IU), which carries the prosodic nucleus of the illocution, and is delimited by a terminal prosodic break. Other types of prosodic units, such as prefix (Topic), suffix (Appendix), postfix (Parenthesis), among others (t'Hart et al., 2006), convey different information functions. These functions pertain to the constitution of the semantic and syntactic content of the utterance (textual IUs) or to the regulation of interaction (dialogic IUs, corresponding to what is called "discourse markers" within other frameworks: see Raso and Vieira, 2016 and Gobbo, 2019 for a new classification of dialogic units). In Table 7 (adapted from Moneglia and Raso, 2014, pp. 490-491) the main textual and dialogic IUs according to the L-AcT are presented, together with their tags:

Table 7: Main IUs according to L-AcT (source: Moneglia and Raso, 2014, pp. 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.

An example of the transcript with the informational tagging is shown in (a):

(a) bfamdl03 [11][4]
*LUZ: *porque quando cê chega num lugar que cê se sente em casa* /=TOP= *cê sabe imediatamente* //=COM=
because when you arrive at a place where you feel at home / you know it right away //

[4] Within the C-ORAL family corpora, specific acronyms identify each file, as *bfamdl03* in (a): *b* stands for Brazilian Portuguese (in the other corpora: *i* for 'Italian', *e* for 'Spanish', *f* for 'French'); *fam* stands for 'private-familiar' context (vs. *pub* for 'public'); *dl* stands for 'dialogue' (other interaction typologies are indicated by *mn* 'monologue' and *cv* 'conversation'); the final number, as *03* above, indicates the number of the file within the respective section (in the example, the third dialogue of the private-familiar context); the number in square brackets corresponds to the utterance number.

The transcripts of the minicorpora follow the same criteria of the reference corpora: non-terminal and terminal prosodic breaks are marked by '/' and '//', respectively, and the starred abbreviation stands for the speaker's acronym (details on transcription criteria are found in Mello et al. 2012). Informational tags are simply added to the original transcripts, as shown in example (a).

2.3.2 Beyond DB-IPIC: the C-ORAL-BRASIL minicorpora

The DB-IPIC platform has been a pioneering project for the cross-linguistic study of information structure, because it is endowed with a multi-level query interface that permits complex interrogations (query parameters involve not only information structure, but also PoS, KWIC, utterance typology, interactional typology, among others). The C-ORAL-BRASIL project has been making efforts to widen the scope of the cross-linguistic comparison, not only giving its contribution to the IPIC platform with the Informal Brazilian Portuguese minicorpus, but also compiling and/or preparing new minicorpora, that are briefly presented in the following sections.

2.3.2.1 The American English minicorpus

An informationally tagged minicorpus of spoken American English, comparable to the Italian and Brazilian Portuguese ones, was compiled (Cavalcante and Ramos, 2016) by extracting a selected group of 20 recordings from the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000). The sampling criteria included the highest possible acoustic quality, the greatest possible diaphasic variation, balance between male and female voices, and the aforementioned 2/3 vs. 1/3 proportion between dialogic (dialogues + conversations) and monologic interactions.

As the reference corpus for this resource is not part of the C-ORAL family, a thorough adaptation of the transcripts to the C-ORAL criteria, together with the prosodic breaks' annotation and a new text-to-speech alignment were necessary (metadata from the SBCSAE were adapted to the C-ORAL-BRASIL format as well; see Cavalcante and Ramos, 2016 for details about the whole compilation process).

In Table 8, the Italian, Brazilian Portuguese and American English minicorpora's sizes are compared, in word and utterance number:

Table 8: Comparability of Italian, Brazilian Portuguese, and American English minicorpora

Italian	**Monologues**		**Dialogues**		**Conversations**		**Total**
words	11,818	37.1%	10,409	32.7%	9,623	30.2%	31,850
utterances	1,347	24%	2,303	41%	1,972	35.1%	5,622
Brazilian Portuguese							
words	9,135	32.1%	10,660	37.5%	8,662	30.4%	28,457
utterances	994	18.1%	2,451	44.7%	2,039	37.2%	5,484
American English							
words	9,359	35.4%	10,647	40.2%	6,464	24.4%	26,470
utterances	992	28.7%	1,382	40%	1,078	31.2%	3,452

As Table 8 shows, the Italian and the Brazilian Portuguese minicorpora are closer in size in what concerns word and utterance numbers; the American English minicorpus displays a similar word number, but it has a lower number of reference units. This difference is partly due to the nature of the SBCSAE, which contains less actional interactions (chats between friends and alike), hence, a lower number of speech acts (the utterance corresponds to one speech act, according to the L-AcT). Still, the difference in number of reference units does not affect the overall comparability between the three minicorpora, and the American English minicorpus represents a very important resource for cross-linguistic comparison, since it provides data of a non-Romance language. This minicorpus is available for download at www.c-oral-brasil.org > Corpora.

2.3.2.2 Brazilian Portuguese new minicorpora

Besides the American English minicorpus, new minicorpora of Brazilian Portuguese have been compiled or will be compiled soon.

First, a new, revised version of the Informal minicorpus has been made available (it can be downloaded at the www.c-oral-brasil.org > Corpora). The main target of the revision process was the informational tagging, but prosodic breaks' segmentation and other transcription criteria were revised as well (Cavalcante et al. forthcoming). The revised BP minicorpus is of special interest also because new tags for dialogic units have been introduced, as a consequence of a thorough study of dialogic units based on strictly prosodic parameters (see Raso and Vieira, 2016; Gobbo, 2019 for details).

Besides the new Informal minicorpus, a Telephone minicorpus, extracted from the Telephone corpus described in 2.2.3, will be soon available for download at www.c-oral-brasil.org > Corpora. This resource's architecture was built to be comparable to the dialogues of the Informal corpus, and it represents a useful tool to study dialogic interaction in different diamesic dimensions (face-to-face vs. telephonic). A comparison between the Telephone minicorpus and the dialogues' section of the Informal corpus is shown in Table 9:

Table 9: Telephone minicorpus vs. Dialogues in C-ORAL-BRASIL Informal corpus

	Telephone minicorpus	**Informal dialogues**
.wav files	27	46
words	8,667	73,358
utterances	1,729	13,978

An almost perfect balance between the number of words pronounces by male (48.4%) and female (51.6%) speakers is found in the Telephone minicorpus. Informationally tagged minicorpora representative of the Formal and Media corpora will be compiled as well.

2.3.2.3 Minicorpora for other languages

Besides all the new BP minicorpora, other resources will be compiled soon for other languages: a new Informal Italian minicorpus (Bossaglia and Raso, *forthcoming*) compiled through a different selection of texts from the Informal Italian C-ORAL-ROM corpus, and an Angolan Portuguese minicorpus (Rocha, Mello and Raso, 2018).

The new Informal corpus architecture is maintained the same as the BP Informal one in what concerns private-familiar and public contexts, and the proportions between dialogic and monologic interactions. Its main novelty concerns acoustic quality, which is in general higher than the IPIC Italian minicorpus: only one file with C quality was included, and its maintenance

was due to its highly actional, hence, informationally interesting characteristics (the recorded situation involves a man teaching his daughter to drive a car). Besides this new selection of texts pursuing the highest possible acoustic quality, a new informational tagging and a review of the prosodic breaks' segmentation will be executed.

The Angolan Portuguese minicorpus will be compiled thanks to a collaboration between the LEEL lab and the *Projeto Libolo*, coordinated by Carlos Figueiredo (Macau University) and Márcia Santos de Oliveira (São Paulo State University, Brazil). In July 2018, a team from the LEEL lab joined the *Projeto Libolo* in the Libolo community (Southern Kwanza region of Angola), recording a diverse set of spontaneous interactions (see Rocha, Mello and Raso 2019 for details on the C-ORAL-ANGOLA minicorpus). Transcription, segmentation, informational tagging, text-to-speech alignment and revisions of the recorded sessions will take place at the LEEL lab. This resource will be of great interest not only because it will add a new diatopic variety of spoken Portuguese (besides the Brazilian one, and the European in C-ORAL-ROM), but also because it could be the first spoken corpus of Angolan Portuguese.

3 Conclusions

The C-ORAL-BRASIL project has now completed the compilation processes for the four Brazilian Portuguese corpora (Informal, Formal, Media, Telephone), which will represent the out-of-Europe branch of the C-ORAL family of spoken corpora.

Besides these resources, the project has committed to the enlargement of the set of informationally tagged minicorpora, not only for Brazilian Portuguese, but also for other diatopic varieties of Portuguese (the C-ORAL-ANGOLA minicorpus), and other languages, within and outside the Romance domain (the new Italian minicorpus, the American English minicorpus).

Hopefully, the C-ORAL-BRASIL resources will contribute to many (cross-linguistic) studies on spoken language in many levels.

REFERENCES

1. Austin J. L. 1962. How to do things with words. The William James. 1978.

2. Barbosa PA, Raso T. Spontaneous Speech Segmentation: Functional and Prosodic Aspects with Applications for Automatic Segmentation/A segmentação da fala espontânea: aspectos prosódicos, funcionais e aplicações para a tecnologia. Revista de Estudos da Linguagem. 2018 Oct 1;26(4):1361-96.

3. Bick E. The parsing system Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000.

4. Boersma P, Weenink D. PRAAT: doing phonetics by computer (5.0. 21). From http://www. fon. hum. uva. nl/praat. 2016.

5. Bossaglia G, Raso T. The C-ORAL-BRASIL minicorpus of spoken Italian. Forthcoming.

6. Cavalcante FA, Ramos AC. The American English spontaneous speech minicorpus. Architecture and comparability. CHIMERA: Romance Corpora and Linguistic Studies. 2016;3(2):99-124.

7. Cresti E. Corpus di italiano parlato. 1. Introduzione. Presso L'Accad. della Crusca; 2000.

8. Cresti E, Moneglia M, editors. C-ORAL-ROM: integrated reference corpora for spoken Romance languages. John Benjamins Publishing; 2005 May 9.

9. Du Bois JW, Chafe WL, Meyer C, Thompson SA, Martey N. Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium. 2000.

10. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971 Nov; 76(5):378.

11. Gobbo O. Marcadores discursivos como unidades informacionais prosodicamente marcadas. 2019 (MA dissertation, UFMG).

12. Hart JT, Collier R, Cohen A. A perceptual study of intonation: an experimental-phonetic approach to speech melody. Cambridge University Press; 2006 Nov 23.

13. Martin P. WinPitch. Pitch Instruments Inc. 2003.

14. Mello H, Methodological issues for spontaneous speech corpora compilation. The case of C-ORAL-BRASIL. Spoken Corpora and Linguistic Studies. 2014 Nov 14: 27-68.

15. Mello H, Raso T, Mittmann M, Vale H, Côrtes P. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG. 2012:125-76.

16. Moneglia M, Martin Ph. The C-ORAL-ROM resource. In C-ORAL-ROM: integrated reference corpora for spoken Romance languages. John Benjamins Publishing; 2005 May 9:1-70.

17. Moneglia M, Raso T. Notes on Language into Act Theory. Spoken corpora and linguistics studies. Amsterdam/New York, Benjamins. 2014:468-89.

18. Nicolas Martinez C, Lombán M. Mini-Corpus del español para DB-IPIC. CHIMERA. Romance Corpora and Linguistic Studies. In press.

19. Panunzi A, Gregori L. DB-IPIC. An XML database for the representation of information structure in spoken language. InPragmatics and Prosody 2011 (pp. 133-150). Firenze University Press.

20. Panunzi A, Mittmann MM. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. Spoken Corpora and Linguistic Studies. 2014 Nov 14:189-227.

21. Raso, T. O corpus C-ORAL-BRASIL. In Raso, T, Mello, H editors. C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal. I. pp. 55-90. Editora UFMG; 2012.

22. Raso T, Mello H, editors. C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal. I. Editora UFMG; 2012.

23. Raso T, Mello H. The C-ORAL-BRASIL I: reference corpus for Informal spoken Brazilian Portuguese. In International Conference on Computational Processing of the Portuguese Language 2012 Apr 17 (pp. 362-367). Springer, Berlin, Heidelberg.

24. Raso T, Vieira MA. A description of Dialogic Units/Discourse Markers in spontaneous speech corpora based on phonetic parameters. CHIMERA: Romance Corpora and Linguistic Studies. 2016;3(2):221-49.

25. Rocha B, Mello H, Raso T. Para a compilação do C-ORAL-ANGOLA: um corpus de fala espontânea informal do português angolano. Filologia E Linguística Portuguesa. 2018 Dec 30;20(Especial):139-57.

26. Santos SM, Raso T, Manual validation of transcription criteria of the C-Oral-Brazil II language resource: assessed criteria, methodology, and results. Forthcoming.

27. Schiel F, The validation of speech corpora. 2004.

28. Teixeira B, Barbosa P, Raso T. Automatic Detection of Prosodic Boundaries in Brazilian Portuguese Spontaneous Speech. In International Conference on Computational Processing of the Portuguese Language 2018 Sep 24 (pp. 429-437). Springer, Cham.

29. van den Heuvel H, Iskra D, Sanders E, de Vriend F. Validation of spoken language resources: an overview of basic aspects. Language Resources and Evaluation. 2008 Mar 1;42(1):41-73.

30. Vieira MA, Raso T, Oliveira E. Métodos automáticos de avaliação da qualidade acústica. Forthcoming.