The American English spontaneous speech minicorpus

Architecture and comparability

Frederico Amorim Cavalcante, Adriana Couto Ramos
Universidade Federal de Minas Gerais

This paper presents the American English (AE) minicorpus, a spontaneous speech resource created within the auspices of the C-ORAL-BRASIL project consisting of texts selected from the *Santa Barbara Corpus of Spoken American English*. We focus on the sampling strategy that guided the selection of texts, the transcription criteria that were implemented and the prosodic and informational annotation carried on the AE minicorpus. The minicorpus was designed to be comparable to the minicorpora of the C-ORAL projects for Italian and Brazilian Portuguese, which were conceived to allow the study of information structure in spontaneous speech in accordance with the principles of the *Language into Act Theory*. This theory comprises a pragmatic framework for the study of spontaneous speech and it integrates the IPO approach into its prosodic model. The IPO approach consists of a perception-based model for the study of intonation, providing an apparatus for the description and classification of melodic contours observed in spontaneous speech.

Keywords: spontaneous speech, Language into Act Theory, information structure, corpus annotation

1. Introduction

In this paper we present the American English (AE) minicorpus (Cavalcante 2016; Ramos 2015), the first linguistic resource of spontaneous AE speech created in accordance with the methodology for speech corpora compilation and annotation developed within the framework of the *Language into Act Theory* (L-AcT; Cresti 2000). The AE minicorpus consists of 20 carefully selected texts from *The Santa Barbara Corpus of Spoken American English* (SBCSAE; Du

CHIMERA. Romance Corpora and Linguistic Studies 3.2 (2016), 99-124. ISSN 2386-2629

Bois *et al.* 2000-2005), and was conceived to be comparable to the C-ORAL family minicorpora of Brazilian Portuguese (BP) and Italian (IT) (DB-IPIC).

In what follows, we discuss some of the issues involved in the creation of the AE minicorpus. First, we present the C-ORAL family, its main corpora and its theoretical and methodological apparatus. Then, we present the C-ORAL family minicorpora, whose architecture served as the model for the AE minicorpus. Finally, we present the main features of the AE minicorpus.

2. The C-ORAL family

What we refer to here as the *C-ORAL family* comprises an international cooperation for speech corpora compilation and spontaneous speech studies. The C-ORAL family first began with the C-ORAL-ROM project (Cresti & Moneglia 2005), which provides the scientific community and speech industry with a set of comparable spontaneous-speech corpora of the four main Romance languages, i.e. French, Italian, European Portuguese (EU) and Spanish. The C-ORAL-ROM is headed by E. Cresti and M. Moneglia at the LABLITA[1] lab (University of Florence). Joining the C-ORAL-ROM as its fifth branch, the C-ORAL-BRASIL (Raso & Mello 2012) is a project dedicated – albeit not exclusively – to Brazilian Portuguese spontaneous speech. It is headed by T. Raso and H. Mello at the LEEL[2] lab (Federal University of Minas Gerais).

The main achievements of the C-ORAL family with respect to corpora compilation is the above-mentioned C-ORAL-ROM multilingual corpus and the C-ORAL-BRASIL corpus of spontaneous Brazilian Portuguese speech. These are comparable speech corpora whose design focus on the representation of varied communicative situations based on the fact that the relative frequency of speech act types correlates with the type of communicative situation (Moneglia 2011) – as opposed to the frequency of lexical items, which vary in accordance with the topic of the conversation. Therefore, the C-ORAL family corpora capture the widest possible range of communicative situations, which, in

[1] *Laboratorio Linguistico del Dipartimento di Italianistca dell'Università di Firenze* (Research Unit at the Humanities Department of the University of Florence): http://lablita.dit.unifi.it/.

[2] *Laboratório de Estudos Empíricos e Experimentais da Linguagem* (Laboratory of Empirical and Experimental Language Studies): www.letras.ufmg.br/leel/.

Weinreich's (1954) diasystem (see also Berruto 1987)[3], is referred to as diaphasic variation.

The C-ORAL family corpora are structured into different sessions. Regarding their informal parts[4], they are subdivided into three interactional typologies (monologues, dialogues and conversations) and two sociological contexts (family/private and public domains). Regarding the size of the corpora, each component of the informal part of the C-ORAL-ROM has, on average, 160,400 words, while the corresponding part in the C-ORAL-BRASIL has 208,130 words.

These corpora were compiled and organized following a common design, thus enabling crosslinguistic studies. They bring prosodic annotation of tone unit boundaries and text-to-speech alignment at the level of utterances, in accordance with the theoretical principles of the framework within which their design was conceived (Cresti 2000).

2.1 The L-AcT framework

Within L-AcT, speech is conceived of as the result of pragmatic activities by the speaker. Thus, the reference unit for speech behavior is associated with the performance of a speech act (Austin 1962). The *utterance* is defined as the smallest perceptually-detectable linguistic unit in the speech stream, showing pragmatic and prosodic autonomy (Cresti 2000). In other words, the utterance is the linguistic counterpart of an act, and is necessarily signaled by a terminal (conclusive) prosodic break.

The annotation of prosodic boundaries conducted on the C-ORAL family corpora is based on the perceptual (auditory) relevance of prosodic cues. As already mentioned, the completion of an utterance is signaled by a prosodic break perceived as terminal. When utterances are realized in more than one tone unit, their internal prosodic boundaries are perceived as non-terminal. The prosodic annotation scheme adopted in C-ORAL family corpora (Moneglia & Cresti 1997) use double-slash signs ("//") to signal terminal prosodic breaks and one-slash signs ("/") to signal non-terminal ones. Other symbols used are the plus sign ("+"), which indicates interrupted utterances and a combination of a one-slash sign and a number enclosed within square brackets ("[/n]"), which indi-

[3] The other variables of the diasystem are: diatopic variation (geographical origin), diastractic variation (sociolinguistic variables: sex, age, schooling, etc.), diamesic variation (medium and channel of a language modality).

[4] For a description of the C-ORAL-ROM formal part, see Cresti & Moneglia (2005). The formal part of the C-ORAL-BRASIL is yet to be finished and published, but it is at an advanced stage of compilation.

cates retracting phenomena. Table 1 below provides the main tags used for the prosodic annotation of the C-ORAL corpora. The tags "…" and "?", which respectively indicate (i) intentionally *suspended* utterances and (ii) terminal breaks with clear interrogative contours, are exclusively used in the C-ORAL-ROM. The C-ORAL-BRASIL maintains the double-slash sign in both cases.

Table 1. Prosodic annotation scheme used in the C-ORAL family corpora

Symbol	Value
//	Indicates a terminal break, marking all prosodically autonomous sequences that do not belong to the previous classes.
/	Signals non-terminal prosodic breaks.
[/n]	Represents retracting phenomena (i.e. false starts), where "n" corresponds to the number of retracted words. Retracting marks can be considered a type of non-terminal break, but the words in false starts do not contribute to the informational patterning nor to the semantic content of the utterance.
+	Signals unintentionally interrupted sequences. In this case, the speaker's program is broken and the interpretability of the sequence can be compromised.
?	Delimits a prosodically autonomous sequence with a clear interrogative prosodic profile.*
…	Delimits a prosodically autonomous sequence voluntarily interrupted by the speaker with a *suspended* prosodic profile.*

* Symbols only used in the C-ORAL-ROM corpus. In the C-ORAL-BRASIL corpus the double-slash symbol is used instead.

2.1.1 *The prosodic model*

A central aspect of the L-AcT framework is the hypothesis that establishes a correspondence between the units of the prosodic pattern (tone units) and those of the informational pattern (information units). This is known as the *Informational Patterning Hypothesis* (Cresti & Moneglia 2010), and comprises an integration of the IPO approach for the perceptual study of intonation (see 't Hart *et al*. 1990) to the pragmatic orientation of the theory.

The IPO model is based on the perceptual relevance of prosodic parameters, particularly that of fundamental frequency (*f0*) variation. The model establishes a correlation between voluntary production of *f0* change (pitch movements) on the part of the speaker and perceptual relevance on the part of the listener. In very general terms, the model proposes that pitch movements are combined into configurations, and that these arrangements of pitch movements are what makes up the melodic contour of utterances in spoken language. Since utterances may

be made up of only one pitch movement, a configuration may consist of one movement alone.

The model also establishes a hierarchy of configuration types. Thus, there are the *root configuration*, whose main feature is that of being obligatory in every contour; the *prefix configuration*, which is optional and necessarily precedes the root; and the *suffix*, which is also optional but necessarily follows the root. The hierarchy is formalized as *(Prefix)^n Root (Suffix)*, where parentheses indicate optionality and the superscripted "n" indicates possibility for iteration.

Regarding the aforementioned correspondence between the units of the prosodic informational patterns, the L-AcT framework associates the root configuration (tone unit) of the IPO model with the realization of a speech act, a pragmatic function that is performed by the *comment* information unit (Cresti 2000). The comment, as the informational counterpart of the root unit, is the only information unit that is necessary and sufficient for the realization of an utterance.

The units of the information pattern are defined in terms of (i) pragmatic function, (ii) prosodic features and (iii) distribution (position) of the unit within the hosting utterance with respect to the comment. Therefore, within L-AcT, the pragmatic functions fulfilled by tone/information units are associated with specific prosodic characteristics. This association comprises the core of the Informational Patterning Hypothesis. Table 2 below shows the correspondence between the prosodic and informational patterns.

Table 2. Relation between prosodic and information patterns

Prosodic Pattern				**Information Pattern**	
	Root		→	Comment *Tag: COM*	
(prefix)		(suffix)	→	(Topic) *Tag: TOP*	(Appendix) *Tag: APC, APT*
	(introducer)		→	(Locutive Introducer) *Tag: INT*	
	(parenthetical)		→	(Parenthetic) *Tag: PAR*	
(incipit)		(phatic)	→	(Incipit) *Tag: INP*	(Phatic) *Tag: PHA*

Parentheses indicate optionality

The annotation of information functions is a prerequisite for the study of information structure within the L-AcT framework. However, the identification and

annotation of information units in a speech corpus is an endeavor that requires both time and human resources, since it is has to be done manually in a process that involves careful examination of each tone unit. Given the size of the C-ORAL family corpora, complete informational annotation on them would be a rather difficult task. Hence, the C-ORAL minicorpora of Italian and Brazilian Portuguese (henceforth IT and BP minicorpora, respectively) were created (DB-IPIC). In must be noted that the size of the minicorpora suffice for the purposes to which they were created, namely the study of the linear relations among information units.

The C-ORAL family minicorpora are, as it were, the models after which the AE minicorpus was created. In the next section, we briefly present the two minicorpora.

2.2 The C-ORAL family minicorpora

The C-ORAL family minicorpora comprise 20 carefully selected texts from the informal sessions of the C-ORAL-ROM (Italian component) and the C-ORAL-BRASIL speech corpora, whose overall architecture is reproduced in the minicorpora. Thus, the minicorpora are structured into balanced sessions in accordance with the interactional typologies documented in the matrix corpora. Also, the minicorpora contain texts from both family/private and public sociological contexts. As previously mentioned, the minicorpora were created in order to allow crosslinguistic studies on the linear relations among information units (see Mittmann & Raso 2011; Panunzi & Mittmann 2014). Thus, besides the annotation of prosodic boundaries, the minicorpora feature annotation of information units.

The sampling criteria adopted for their creation are summarized below:

- Selection of texts showing good *acoustic quality*, in order to allow appropriate assessment of prosodic parameters;
- Search for *diaphasic variation*, in order to capture the greatest possible range of illocutionary types (see Moneglia 2011);
- Search for *equilibrium of male and female voices* in the samples, since prosodic parameters correlate with speaker's sex;
- Search for a *balanced number of words in the interactional typologies* – 1/3 in monologues and 2/3 in dialogues and conversations together[5];

[5] Dialogues and conversations share many common features. For instance, both these interactional typologies are highly context-oriented and tend to be centered on speech-act performance rather than textual construction, as opposed to monologues. The main difference be-

- Search for texts with *content of interest to annotators*, so as to reduce the amount of erros due to inattention.

Table 3 provides a description of each of the texts of the minicorpora, thus showing their diaphasic variation.

Table 3. IT and BP minicorpora texts

	Italian minicorpus	**Brazilian minicorpus**
Monologues	01 Interview with an old partisan at his home	01 Man tells a story about a snake
	02 Elderly woman tells life story to her relatives	02 Grandmother tells family stories to grandson
	03 Narrative to a relative about the honeymoon	03 Father tells family two entertaining stories
	04 An after-dinner travel tale to friends	04 Woman tells about her experience in the hospital
	05 Interview with a retired traveling salesman	05 Woman shares the story about her daughter's adoption
	06 Political speech at a political-party meeting	06 Man explains his professional trajectory
	07 Professional explanation to a colleague about office work	07 Interview with public school teacher
	08 Interview with an employee of the Poggibonsi municipality	
Dialogues	01 Interview of an artisan in his leather workshop	01 Two friends shop for groceries
	02 Friends at home making a cake	02 Two colleagues chat while packing recording equipment
	03 Beautician and customer in the beauty-center	03 Couple takes a car trip
	04 Two friends develop photos in a darkroom	04 Maids do the dishes

tween them is the number of participants (two in dialogues and more than two in conversations). This is mainly the reason for considering the two types together. For more about the distinction between interactional typologies see Mello (2014).

	05 Father gives driving lesson to his daughter	05 Broker shows apartment to his sister
	06 Proposal of an insurance policy	06 Engineer and construction worker at construction site
	07 Teachers' meeting at the school office	07 Customer and salesman in a shoe store
Conversations	01 Relatives talk while browsing through family photos	01 Young friends evaluate a soccer championship
	02 Friends explain the game Mastermind	02 Elderly ladies chat about an upcoming marriage
	03 Family talks with child during lunch preparation	03 Friends play snooker
	04 Meeting of a voluntary association	04 Friends play Pictionary
	05 Chat at a hardware store while shopping	05 Employees at a blood bank explain their work
		06 Political meeting

The table above offers a glimpse of the diaphasic variation in the IT and BP minicorpora. The monologues, as spoken interactions in which there is *prevalence* of textual elaboration by one speaker, consist of narratives and interviews only, for prototypical monologues, particularly in informal settings, are actually quite rare in spontaneous speech (see Mello 2014). Dialogues and conversations, on the other hand, consist of different types of communicative situations, e.g. sales encounter, verbal interactions while cooking, cleaning, shopping, driving, and also chats, game playing, and work meetings.

Regarding the sociological contexts, the IT minicorpus has 14 family/private texts and 6 public ones, whereas the BP minicorpus has 15 family/private texts and 5 public ones.

All the texts that compose the minicorpora were annotated with tags identifying informational functions. The prosodic boundaries had already been annotated prior to the creation of the minicorpora. The informational annotation phase was conducted in accordance with the L-AcT principles (Cresti 2000) and the Informational Patterning Hypothesis (Cresti & Moneglia 2010).

Information units may be either textual or dialogic. Textual information units make up (or refer to) the very text of the utterance, while dialogic units regulate the interaction and are directed at the interlocutor.

Textual units are divided into illocutionary and non-illocutionary ones. The first group contains the units that carry the illocutionary force, while the non-illocutionary group contains the textual units whose functions are not directly related to speech act performance.

Table 4 shows the tagset used for the annotation of information functions, along with a brief definition for each tag. Note that there are other, non-informational units in the table. For details regarding such units, see Panunzi & Mittmann (2014).

Table 4. Tagset used in the annotation of the C-ORAL family minicorpora

Unit type	Name	Tag	Definition
Textual (illocutionary)	Comment	COM	Carries the illocutionary force of the utterance. It is necessary and sufficient for the performance of the utterance.
	Multiple Comment	CMM	Constitutes a chain of Comments, which form an illocutionary pattern, i.e. an action model which allows the linking of at least two illocutionary acts for the performance of one conventional rhetoric effect.
	Bound Comment	COB	A sequence of Comments, which are produced by progressive adjunctions that follow the flow of thought. It forms a distinct speech unit, the Stanza.
Textual (non-illocutionary)	Topic	TOP	Supplies the domain of application for the illocutionary act, providing a cognitive reference to the speech act. It allows the utterance to be displaced from its immediate context (linguistic and non-linguistic).
	Topic List	TPL	A sequence of two or more (normally three) semantically and syntactically connected units that form only one prosodically marked major unit of Topic.
	Appendix of COM	APC	Integrates the text of the Comment.
	Appendix of TOP	APT	Integrates the text of the Topic.
	Parenthetic	PAR	Inserts information into the utterance with a metalinguistic value; its scope can be backward, forward or both.

	Parenthetic List	PRL	A sequence of two or more (normally three) semantically and syntactically connected units that form just one prosodically marked main unit of Parenthesis.
	Locutive Introducer	INT	Expresses the evidence status of the subsequent locutive space (simple or patterned) marking a shift of the pragmatic coordinates for its interpretation.
Dialogic	Incipit	INP	Opens the communicative channel, bearing a contrastive value. Starts dialogic turns or utterances.
	Conative	CNT	Pushes the interlocutor to do or stop doing something.
	Phatic	PHA	Controls the communicative channel, ensuring its maintenance.
	Allocutive	ALL	Specifies to whom the message is directed and enacts social cohesion.
	Expressive	EXP	Works as an emotional support, stressing the sharing of a social affiliation.
	Discourse Connector	DCT	Connects different parts of the discourse, indicating its continuation.
Non-informative	Scanning Unit	SCA	Used when a Prosodic unit does not bear an information nucleus and does not signal any information function, but rather scan the locutive content.
	Interrupted unit	i-[TAG]	For instance: i-COM means that a COM is interrupted by a parenthetic or a dialogic unit and its completion will follow afterwards, e.g. *John said* /=i-COM *or this is what I remember* /=PAR= *that he likes pasta* //=COM=.
	Empty unit	EMP	Used when one prosodic unit is filled with material whose informational content is not to be considered in the overall content of the utterance as happens when: (a) there is a retracting; (b) the last unit of an utterance is interrupted; e.g. *John says* [/2]=EMP= *John said that he likes pasta* //=COM=.
	Time Taking	TMT	Tag used for the so-called filled pauses.

	Unclassified	UNC	Unclassifiable Unit. It is not possible to attribute another tag to the unit for some reason.
Other	Reported unit	[TAG] _r	Indicates that the information is part of a reported speech.

The C-ORAL family minicorpora are available online through the Information Structure Database (DB-IPIC; Panunzi & Mittmann 2014) at the LABLITA website. The platform hosting the database has a user-friendly design and allows, among other things, the study of linear relations among information units both within each corpus as well as across them.

Table 5 below shows the size of the C-ORAL family minicorpora, both in number of words and reference units. The term *reference unit* (RU) refers to utterances and stanzas. *Stanzas* are pragmatically and prosodically autonomous units, but, unlike utterances, they are marked by a tempo that reflects the *unpatterned* production of more than one speech act, each of which rising from a distinct intentionality (see Moneglia & Raso 2014). Comment units in a stanza, except the last one in the sequence, are called *bound comments* (COB; see Table 4), for they are marked by a prosodic signal of continuity.

Table 5. Size of the IT and BP minicorpora[6]

IT minicorpus	**Monologues**		**Dialogues**		**Conversations**		**Total**
Words	11818	37,1%	10409	32,7%	9623	30,2%	31850
Total RU	1347	24,0%	2303	41,0%	1972	35,1%	5622
BP minicorpus	**Monologues**		**Dialogues**		**Conversations**		**Total**
Words	9135	32,1%	10660	37,5%	8662	30,4%	28457
Total RU	994	18,1%	2451	44,7%	2039	37,2%	5484

The IT minicorpus has 31,850 words, while the BP minicorpus has 28,457 words. Nevertheless, this is not deemed an important difference (Panunzi & Mittmann 2014), since the proportion of words in monologues, on the one hand, and dialogues and conversations, on the other, is sufficiently similar. Furthermore, the number of RUs in each minicorpus is very close, which is a more rel-

[6] The statistics provided in Table 5 differ a little from those provided in Panunzi & Mittmann (2014), because we have used an *R* script that does not consider interrupted words, time-taking tokens, indications of retracting and of paralinguistic noise, among other things, for word computations.

evant feature than word counts for minicorpora whose main purpose is the study of how information is structured in spontaneous speech.

Like the C-ORAL corpora, the IT and BP minicorpora feature text-to-speech alignment at the level of utterances, carried with Winpitch (Martin 2005). The alignment process comprises the univocal association of previously determined units in the transcription file and their corresponding portion in the audio file. The utterance-based alignment performed on the C-ORAL corpora reflects the pragmatic orientation of the framework (Cresti 2000; Moneglia & Raso 2014) that guided their creation.

Text-to-speech alignment is of utmost importance for speech studies, since, without it, it is hardly possible to locate in the acoustic signal specific parts of transcriptions. Moreover, an unaligned speech corpus is likely to favor a methodology that takes the transcription – in itself a limited representation of speech – for the object that the transcription is meant to represent (Mello 2014).

In the next section we will present the AE minicorpus, which was created following the same parameters adopted in the creation of the IT and BP minicorpora.

3. The AE minicorpus

The AE minicorpus (Ramos 2015; Cavalcante 2016) is a set of 20 carefully selected texts from the SBCSAE (Du Bois *et al*. 2000-2005). It was created at the LEEL lab, under the guidance of Prof. T. Raso (Federal University of Minas Gerais), and comprises the first linguistic resource of a non-Romance language created within the pragmatic framework of the L-AcT.

In order to ensure comparability with the C-ORAL family minicorpora, we adopted the same sampling criteria for the AE minicorpus (see section 2.2). Thus, its texts were chosen based on acoustic quality, diaphasic variation, equilibrium of male and female voices, a balanced number of words in the interactional typologies, and content of interest.

After the selection of texts, we implemented the transcription criteria adopted for the C-ORAL-BRASIL corpus (Mello *et al.* 2012) and, concomitantly, performed prosodic annotation on the entirety of the minicorpus. Transcriptions and prosodic annotation were done following a version of the CHAT system (Macwhinney 2000) implemented for prosodic-boundary annotation (Moneglia & Cresti 1997). The tagset used for the annotation of prosodic boundaries on the AE minicorpus is the same used on the C-ORAL-BRASIL corpus (see Table 1).

The utterances identified during the annotation phase and the corresponding acoustic sources were aligned using the software Winpitch (Martin 2005). The text-to-speech alignment was carried at the level of utterances, in accordance with the IT and BP minicorpora. Finally, the AE minicorpus received annotation of informational functions (for the tagset used, see Table 4).
Before providing more details about the minicorpus itself, we will briefly present the SBCSAE, its matrix corpora.

3.1 Matrix corpus

The Santa Barbara Corpus of Spoken American English, from which the AE minicorpus was created, is a corpus of spontaneous American English speech collected by researchers at the Center for the Study of Discourse at the University of California, Santa Barbara (USCB), under the direction of J.W. Du Bois. The corpus contains 60 texts and approximately 249,000 words, documenting formal and informal registers in a variety of spoken interactions, with male and female speakers of different social backgrounds from various locations within the United States.

The features of the SBCSAE that made us consider it as a suitable source of texts to compose the AE minicorpus are summarized below:

- The SBCSAE is a corpus of a non-Romance language of great academic reach;
- It is a *spontaneous* speech corpus, in the sense that it contains speech whose planning and execution takes places in synchronicity (Nencioni 1983);
- It contains audio files with good acoustic quality;
- It documents different communicative situations;
- It is licensed under a Creative Commons attribution[7], which facilitates the access to its content.

The following sections provide details concerning the AE minicorpus.

3.2 Diaphasic variation

One of the major factors that motivated the creation of the AE minicorpus was the fact that the principles and methodology of the L-AcT approach had not yet been taken beyond the boundaries of the Romance languages documented in the C-ORAL family corpora (Moneglia & Raso 2014). In addition, the AE minicor-

[7] Attribution-NoDerivs 3.0 United States (CC BY-ND 3.0 US).

pus was conceived as a resource to facilitate the communication between researchers working within L-AcT and researchers working within other frameworks, since it provides examples of theoretical constructs that were first conceived based on Romance language data in a language of great academic reach. Like the IT and BP minicorpora, the AE minicorpus has 20 texts, divided into monologues, dialogues and conversations. Table 6 presents the AE minicorpus texts in their respective interactional typologies, along with a brief description of each of them[8].

Table 6. Texts in the AE minicorpus (* Public interactions)

Monologues	01 A student explains her studies in equine science in the living room of a house trailer
	02 Two friends/co-workers talk about their interests at work
	03 Two cousins chat at home after a long time apart
	04 A man talks about his experiences as a gay man at home
	05 Two friends talk as they watch TV at home
	06 Two male friends chat about science and human nature at home
	07 A woman talks about penguins at a meeting at an aquarium*
Dialogues	01 Two cousins chat at home
	02 A couple lying in bed talk about a book
	03 Mother and daughter at home talk after work
	04 A man and a woman talk on a visit to her ranch
	05 A couple plays Hearts in a summer house
	06 A work conversation at an air traffic control tower between an experienced air traffic controller and an unexperienced one*
	07 A homeowner and an engineer talking at home about air-conditioning systems*
	08 A salesman and a female buyer at a store discuss different types of tape decks*
Conversations	01 Three friends chat about traveling, health and vitamins in the living room.
	02 Two sisters and their mother talk in a restaurant as they decide on what to eat
	03 Friends talk at a block party
	04 Family members chat at a birthday party
	05 Friends talk at a dinner party

[8] For word counts see Section 3.5.

Most of the texts of the AE minicorpus (16) belong to the family/private sociological context, and four texts (one monologue and three dialogues) belong to the public context. Therefore, as Panunzi & Mittmann (2014) point out regarding the small number of texts produced in the public context in the IT and BP minicorpora, public context cannot be considered a variable proper for studies based on the AE minicorpus, given the limited representation of the context in the minicorpus. In other words, the AE minicorpus is balanced in terms of sociological contexts, but its size is not large enough for the public context to be considered representative, and the same applies to the IT and BP minicorpora.

The criteria used for the classification of interactions as family/private or public are the ones adopted in the C-ORAL-BRASIL project (see Mello 2014). Thus, regardless of where the interaction took place, a variable that was considered in the classification of the C-ORAL-ROM corpora, when speakers performed professional or institutional roles, the text was classified as public; when speakers performed "individual" roles, on the other hand, the text was classified as family/private.

Diaphasic variation is considered of paramount importance for a corpus designed for the study of information structure. As Moneglia (2011) and Mello (2014) consistently argue, in order for a corpus to capture the widest possible range of illocutionary variation, it has to document the widest possible range of communicative situations. Therefore, we favored interactions not merely with different speakers talking about different things, but rather interactions involving the performance of activities besides the verbal interaction (see, e.g., dialogues 04, 05, 07, and 08 and conversation 02 in Table 6).

Nevertheless, the texts in the SBCSAE are, in general, less "actional" than those in the earlier C-ORAL resources, in the sense that many of interactions documented in the SBCSAE are "restricted" to verbal exchanges only, as speakers often are not engaged in the performance of any activity other than the verbal interaction itself. In order to cope with such circumstance, we selected the most interactive extracts from each of the eligible texts of the matrix corpus[9]. The result is a minicorpus slightly less "actional" than the IT and BP minicorpus, but not so much as to compromise the comparability.

[9] All texts in the SBCSAE exceed the average of 1,500 words that comprised our target.

3.3 Acoustic quality

Another criterion considered for the selection of texts to compose the AE minicorpus was the *acoustic quality*. The assessment of acoustic quality was conducted following the criteria provided in Table 7.

Table 7. Criteria for assessment of acoustic quality (adapted from Raso 2012)

Quality	Description
A	Very 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.

The texts that make up the AE minicorpus were acoustically classified as follows:

- Quality A: 1 text;
- Quality AB: 3 texts;
- Quality B: 10 texts;
- Quality BC: 2 texts;
- Quality C: 4 texts.

In the AE minicorpus, 70% of recordings show medium, high or very high acoustic quality. The remaining 30% show either mid-low or low quality. The fact that the SBCSAE was compiled mostly in the late 1980's and early 1990's accounts for the imperfect acoustic quality of some of the recordings in the AE minicorpus, since the recording equipment available at that time was not as sharp as what is now easily available.

3.4 Diastratic profile

In the AE minicorpus, 56% of speakers are female and 44% are male. More significant, however, is the fact that 51.4% of words are uttered by female speakers and 48.6% by male ones, for it is a more reliable indication of the desired equilibrium between male and female voices.
Following the C-ORAL-BRASIL standard, speakers in the AE minicorpus were classified according to the age and schooling groups defined in Table 8 below.

Table 8. Age and schooling groupings adopted for the AE minicorpus

Age	**Schooling**
A 18 to 25 years old	**1** Incomplete basic level or up to 7 years of schooling
B 26 to 40 years old	**2** Up to undergraduate degree as long as not having a profession related to university degree
C 40 to 60 years old	**3** Professions dependent on a university degree
D over 60 years old	**X** Unknown
X Unknown	

In the AE minicorpus, 46.3% of words are uttered by speakers between 26 and 40 years of age (group B), 18.2% by speakers between 18 and 25 (group A), 14.7% by speakers between 40 and 60 (group C), and 12.9% by speakers who were at the time of recording over 60 years of age (group D). Information about the age of some speakers could not be retrieved; therefore, 7.9% of words of the minicorpus are uttered by speakers who fall within the unknown age group.

Regarding schooling, 62.2% of words are uttered by speakers who have up to undergraduate degree[10] (group 2), 24.9% of words by speakers who have university degrees and work in their degree area (group 3), while the schooling of of speakers uttering 12.9% of words is unknown (group X). No participant in the AE minicorpus belongs to the group 1.

The AE minicorpus is therefore more representative of age group B and schooling group 2. In other words, the diastratic profile of the minicorpus cannot be considered perfectly balanced. Nevertheless, a perfect diastratic profile was never sought after, since our methodological choices led us to favor other parameters, mainly diaphasic variation and acoustic quality.

[10] Note that, in order to be included in group 2, speakers with an undergraduate degree must not work in their degree area (see Table 8).

3.5 Size

The twenty texts of the AE minicorpus total 26,470 words, an average of 1,300 words per text. Thirteen texts have 1,000 to 1,500 words, while four texts have less than 1,000 words (two monologues with 567 and 344 words, and 2 conversations with 954 and 855 words) and three texts have more than 1,500 (two monologues of 1,708 and 2,566 words, and one conversation of 2,050 words). In favoring acoustic quality, speech event structure (monologue, dialogue, conversation) and diaphasic variation, we could not possibly achieve a perfectly uniform minicorpus with respect to number of words in each text.

Regarding RUs in the AE minicorpus, their proportions in each interactional typology is very similar to what we see in the IT and BP minicorpora:

- RUs in the *IT minicorpus*: 24% in monologues and 76% in the dialogical typology (dialogues and conversation);
- RUs in the *BP minicorpus*: 18% in monologues 82% in the dialogical typology;
- RUs in the *AE minicorpus*: 29% in monologues and 71% in the dialogical typology.

The distribution of words and RUs in each interactional typology in the AE minicorpus is shown on Table 9.

Table 9. Size of the AE minicorpus – words and RUs

	Monologues		**Dialogues**		**Conversations**		**Total**
Words	9359	35,4%	10647	40,2%	6464	24,4%	26470
RUs	992	28,7%	1382	40,0%	1078	31,2%	3452
simple RUs	450	24,0%	774	41,3%	650	34,7%	1874
compound RUs	542	34,3%	608	38,5%	428	27,1%	1578

As the above table shows, the AE minicorpus has 35,4% of its words in monologues and 64.6% in dialogues and conversations considered together. That is to say, the desired balance in terms of proportion of words within interactional typologies (2/3 in monological and 2/3 in dialogical) was attained.

Regarding the distribution of RUs, 24.0% of simple ones (i.e. RUs made up of only one information unit) are in monologues and the remaining 76.0% are in dialogues and conversations. As for compound RUs, 34.3% are in monologues and 65.7% in dialogues and conversations. Dialogues and conversations in the AE minicorpus are, as already discussed, less interactional and therefore less ac-

tional than the same interactional typologies in the IT and BP minicorpus. And this is probably why such a high proportion of compound RUs are found in dialogues and conversations in the AE minicorpus.

3.6 Transcription criteria

The transcription criteria used in the AE minicorpus followed the C-ORAL-BRASIL guidelines (Mello *et al.* 2012), which were designed, among other things, to ensure faithfulness to the recorded content and to preserve readability of texts. Alterations of the original SBCSAE transcription were kept to a minimum, being mostly related to the representation of non-linguistic aspects like shown in Table 10[11].

All the non-standard forms used in the AE minicorpus transcriptions are documented in the header files (see section 3.8) that accompany each text. Paralinguistic noises, hesitations, interrupted words, unintelligible words or sequences, retracting phenomena were represented using the symbols in Table 10 below. The symbols are also part of the C-ORAL-BRASIL guidelines.

Table 10. Symbles used in the AE minicorpus transcriptions

Symbol	Value
hhh	Paralinguistic noise, e.g. laughs, coughs and throat clearings.
&he	Hesitation or time-taking vocalization.
&	Interrupted word; the "&" sign is put immediately before the interrupted word.
< >	Overlapped sequence.
yyy	Anonymized person, institution, telephone number, etc.
xxx	One incomprehensible word.
yyyy	More than one incomprehensible words.

Also following the C-ORAL-BRASIL guidelines, the representation of alphabet letters, acronyms and initialisms was done as follows. Alphabet letters were transcribed as syllables, e.g. letter "a" was transcribed as "ey", "b" as "bee", "c" as "cee", and so forth. Acronyms, on the other hand, were transcribed as a sequence of capital letters. As for initialisms, they were transcribed as a sequence

[11] See below for alterations regarding acronyms and initialisms. For prosodic and informational annotation, see sections 3.7 and 3.9.

of letters, following the convention adopted for the representation of alphabet letters. For example, the acronym MET (*Metropolitan Museum*), pronounced as [mɛt] was transcribed as "MET", while the initialism "VFR" (*Visual Flight Rules*), pronounced as [vi:ɛf'ar], was transcribed as *veeefar*.

3.7 Prosodic annotation

The annotation of prosodic boundaries and the implementation of the transcription criteria of the C-ORAL-BRASIL on the AE minicorpus were conducted in tandem. Given our theoretical choices, the annotation scheme adopted in the SBCSAE, designed within a framework with specificities of its own, could not be repurposed. Thus, after selecting the text extracts from the SBCSAE to compose the AE minicorpus, we removed the symbols and conventions used by the Santa Barbara team from the original transcriptions.

The prosodic annotation process comprises the identification of perceptually relevant prosodic breaks and their evaluation as either terminal or non-terminal (see section 2). Terminal prosodic breaks signal the fulfillment of an utterance, indicating its conclusion, whereas non-terminal prosodic breaks indicate that tone units belong to the same melodic/pragmatic program.

As already mentioned, the double-slash sign ("//") is used for the annotation of terminal breaks, and the single-slash sign ("/") for non-terminal ones. In addition, we use the plus sign ("+") to annotate interrupted utterances, and a slash sign together with a number in between square brackets ("[/n]") to annotate retracting phenomena. The number in the annotation scheme for retracting phenomena indicates the number of words cancelled by the speaker. Retracting phenomena may be thought of as a *programming-execution mismatch*, as speakers break an original program for a tone unit and then reformulate it in the next tone unit.

The examples below provide (1) the original transcription of an SBCSAE extract and (2) its counterpart in the AE minicorpus:

(1) LENORE: .. [So you have your] own equipment,
LYNNE: [(H)]
LENORE: but,
LYNNE: (TSK) (H) No.
I don't have my own equipment at all.
... Da=d,
... you know,
has done some of it. (SBC001, 60'73"-67'17")

(2) *LEO: so you have your own equipment / but +
*LYN: hhh no / I don't have my own equipment at all // dad / you know / has done some of it // (afammn01[1-3])[12]

As it can be noted, the original transcription (Du Bois *et al.* 1993) is more granular, as it indicates pauses ("‥" and "…" for short and medium ones), speech overlapping ("[]") with non linguistic content ("(H)", used for inhalations), vowel lengthening ("="), etc. The AE minicorpus version, on the other hand, in the interest of transcription readability, does not annotate those phenomena, which can nonetheless be easily recovered through the aligned files (see section 3.8).

Tone units are put in separate lines in the original transcription, and punctuation marks are used to indicate pitch movements, which in the extract above can be final (".") or continuing (","). In addition, interrupted units are annotated with the same sign that is used to indicate "continuation" – as seen in the first line uttered by Lenore in examples (1) and (2).

During the prosodic annotation phase, we used the Winpitch program both for listening to the audios and examining the visual representation of acoustic cues. Figure 1 below shows a screenshot of the Winpitch spectrogram window.

3.8 Text-to-speech alignment

The text-to-speech alignment followed the implementation of the transcription and prosodic annotation criteria presented in the previous sections. As already mentioned, the alignment process comprises the assignment of temporal indexes to previously determined units in the transcriptions, thus univocally associating them with their corresponding portion in the acoustic file. The unit chosen for the alignment of the AE minicorpus was the utterance, which is consistent with our theoretical and methodological orientation (see section 2).

Winpitch[13] (Martin 2005) was the software used for the alignment. Figure 2 shows a screenshot of an alignment window of the program.

[12] Following the C-ORAL model, the files in the AE minicorpus are named as follows: the first letter identifies the language ("a" for American English), the next three letters identify the sociological context ("fam" for family/private and "pub" for public), the other two letters identify the interactional typologies ("mn" for monologue, "dl" for dialogue, and "cv" for conversation), and the numbers identify the text within interactional types. Therefore, *afammn01* reads American English minicorpus, first family/private monologue. The numbers provided in square brackets indicate the rank of the utterance within the text.

[13] Free 30-day trial available at http://www.winpitch.com/.

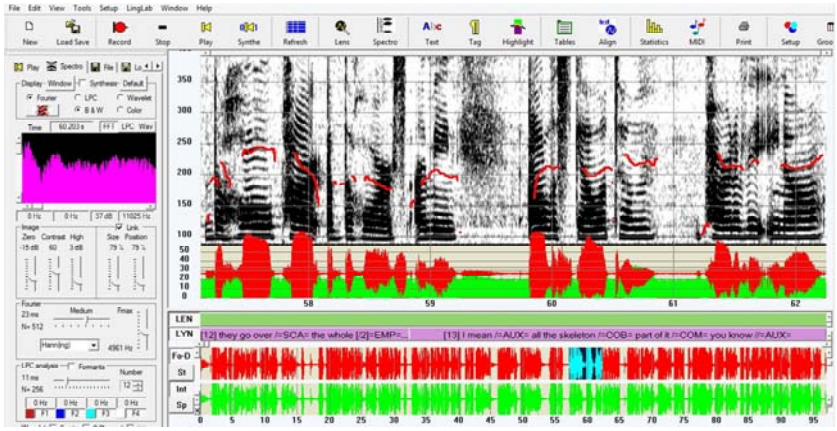

Figure 1. Screenshot of Winpitch spectrogram window

Figure 2. Screenshot of Winpitch alignment window

The aligned texts of the AE minicorpus are sorted into separate folders containing the following files:

- Audio in WAV format;
- Alignment in XML and WP2 formats;
- A document type definition (DTD) file, which is necessary for opening the alignment files in Winpitch;

- Transcript in RTF;
- A header file in TXT.

Table 11 shows an example of a header file of a text from the AE minicorpus:

Table 11. Header file of a text from the AE minicorpus

@Title: Deadly diseases
@File: afamcv01
@Participants: LEN, Lenore, (woman, B, 2, student, participant, Los Angeles/CA), JOA, Joanne (woman, B, 3, teacher, participant, Los Angeles/CA), KEN, Ken, (man, B, 2, photographer/student, participant, Los Angeles/CA)
@Date: 06/02/1987
@Place: private home, residential neighborhood, living room, Los Angeles, California.
@Situation: A conversation among three friends. KEN and JOA are a couple, and LEN is a friend of theirs who is visiting.
@Topic: travel places, vitamins
@Source: Santa Barbara Corpus: SBC015
@Class: informal : particular : conversation
@Length: 5'36"
@Words: 1568
@Acoustic_quality: B
@Transcriber: Adriana Couto Ramos
@Revisor: Adriana Ramos and Frederico Amorim
@Comments: from 1'50'' to 2', speakers laugh. In 5'16'' LEN clears her throat. Apheretic forms: 'till (until)

3.9 Informational annotation

For the annotation of informational values, we used the same tagset that was used for the annotation of the IT and BP minicorpora (see Table 4). The annotation was performed after the alignment phase, since it cannot be done without the simultaneous access to both transcription and acoustic source.

Information units (see section 2) are defined in terms of functional role, prosodic features and distribution within the hosting utterance. The first step for annotating an utterance is to identify the tone unit carrying the illocutionary force. Then, the other units are examined and tagged according to the three criteria used within L-AcT to determine the nature of an information unit, namely func-

tional role, prosodic features and distribution (see section 2). Example 3 below shows the utterances presented in example 2 after the informational tagging.

(3) *LEO: so you have your own equipment /=COB= but + =UNC=
*LYN: hhh no /=CMM= I don't have my own equipment at all //=CMM=
*LYN: dad /=COM= you know /=AUX= has done some of it //=APC=
(afammn01[1-3])

Due to issues related to the irregular acoustic quality of the SBCSAE recordings, dialogic units in the AE minicorpus were frequently tagged as "AUX", which signals their dialogic nature without specifying their function.

Before the informational annotation, the texts of the AE minicorpus had undergone orthographic and prosodic annotation revisions. While in the information annotation phase, which also comprised a revision, we had the opportunity to double check the grammatical and prosodic annotation accuracy of the texts in the AE minicorpus.

4. Conclusion

In this paper we presented the AE minicorpus, the first linguistic resource of a non-Romance language created following the theoretical and methodological principles of the L-AcT framework. The model for the AE minicorpus was the IT and BP minicorpora, which were created from the C-ORAL-ROM and the C-ORAL-BRASIL corpora.

With 26,470 words and approximately 2.5 hours of recordings, the AE comparable minicorpus provides researchers working within the L-AcT framework with valuable means for crosslinguistic comparisons. Moreover, an adapted AE minicorpus expands the possibilities for the dissemination of the L-AcT approach, as its theoretical claims can now be tested and exemplified in a language of great academic reach.

Acknowledgments

We would like to thank M.M. Mittmann for providing the R script that we used for word and reference unit computations; we are also grateful to FAPEMIG for financing this work.

References

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

Berruto, G. 1987. *Sociolinguistica dell'italiano contemporaneo*. Roma: La Nuova Italia Scientifica.

Cavalcante, F. A. 2016. The topic unit in spontaneous American English: a corpus-based study. MA thesis, Federal University of Minas Gerais.

Cresti, E. 2000. *Corpus di Italiano parlato*. Firenze: Accademia della Crusca.

Cresti, E. & Moneglia, M. (eds) 2005. *C-ORAL-ROM: Integrated reference corpora for spoken Romance languages*. Amsterdam: John Benjamins.

Cresti, E. & Moneglia, M. 2010. Informational patterning theory and the corpus-based description of spoken language: The compositionality issue in the topic-comment pattern. In M. Moneglia & A. Panunzi (eds), *Bootstrapping information from corpora in a cross-Linguistic perspective*. Firenze: Firenze University Press, 13-45. http://lablita.dit.unifi.it/publications/bootstrap (accessed January 4, 2016).

DB-IPIC. http://lablita.dit.unifi.it/ipic/ (accessed December 15, 2015).

Du Bois, J.W., Schuetze-Coburn, S., Cumming, S. & Paolino, D. 1993. Outline of discourse transcription. In J.A. Edward & M.D. Lambert (eds), *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Lawrence Erlbaum Associates, 45-89.

Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Englebretson, R. & Martey, N. 2000-2005. *Santa Barbara corpus of spoken American English*, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Martin, P. 2005. WinPitch Corpus: a text to speech analysis and alignment tool. In E. Cresti, & M. Moneglia (eds), *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Amsterdam: John Benjamins, 40-51.

Macwhinney, B. J. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah NJ: Lawrence Erlbaum Associates. http://childes.psy.cmu.edu/manuals/CLAN.pdf (accessed December 15, 2015).

Mello, H. 2014. Methodological issues for spontaneous speech corpora compilation: the case of the C-ORAL-BRASIL. In T. Raso & H. Mello (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 29-68.

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

Mittmann, M. M. & Raso, T. 2011. The C-ORAL-BRASIL informationally tagged minicorpus. In H. Mello, A. Panunzi & T. Raso (eds), *Pragmatics and prosody: illocution, modality, attitude, information patterning and speech annotation*. Firenze: Firenze University Press, 151-183. http://lablita.dit.unifi.it/publications/fup2012_Pragmatics (accessed November 8, 2015).

Moneglia, M. 2011. Spoken Corpora and Pragmatics. *Revista Brasileira de Linguística Aplicada* 11(2): 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*. Pisa: Del Cerro, 57-90.

Moneglia, M. & Raso, T. 2014. Notes on Language into Act Theory. In T. Raso & H. Mello (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 469-495.

Nencioni, G. 1983. Parlato-parlato, parlato-scritto, parlato-recitato. In Nencioni, G. (ed.), *Di scritto e di parlato*. Bologna: Zanichelli, 126-179.

Panunzi, A. & Mittmann, M. M. 2014. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In T. Raso & H. Mello (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 129-151.

Ramos, A. C. 2015. The use of certainty adverbs in Brazilian Portuguese and American English: a semantic/pragmatic approach. MA thesis, Federal University of Minas Gerais.

Raso, T. 2012. O C-ORAL-BRASIL e a Teoria da Língua em Ato. In T. Raso & H. Mello (eds), *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: UFMG, 91-123.

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

SBCSAE. http://www.linguistics.ucsb.edu/research/santa-barbara-corpus (accessed December 15, 2015).

't Hart, J., Collier, R. & Cohen, A. 1990. *A perceptual study on intonation: an experimental approach to speech melody*. Cambridge: Cambridge University Press.

Weinreich, U. 1954. Is a structural dialectology possible? *Word* 10: 388-400.