THE C-ORAL-BRASIL INFORMATIONALLY TAGGED MINI-CORPUS[1]

Maryualê M. Mittmann, Tommaso Raso
Universidade Federal de Minas Gerais (Brazil)

1. Introduction

This paper has two main goals:

- To present a corpus of small proportions that constitutes a sample extracted from the C-ORAL-BRASIL corpus (Raso & Mello 2010; Raso & Mello in press) for spontaneous spoken Brazilian Portuguese. This corpus was tagged with respect to the informational structure following the Language into Act Theory (Cresti 2000) and therefore allows some first consideration about the information structure of Brazilian Portuguese. A comparable mini-corpus was selected for Italian from the Italian C-ORAL-ROM. In the paper we give some first results of the comparation of the two mini-corpora.
- To discuss some interesting aspects of the prosodic annotation of the C-ORAL-BRASIL corpus observing the corrections of the annotation done during the information tagging. The informational tagging is a different perspective from that of the prosodic annotation, and the study of the corrections of the prosodic annotation during the process of informational tagging is useful for better understanding both the perceptual aspects of the prosodic annotation and the cognitive aspects of the informational tagging.

The Brazilian sample is referred as Brazilian mini-corpus, and is a 15 percent (in number of words) portion of C-ORAL-BRASIL informal section[2]. The Italian

[1] This research was financed by CNPq and Fapemig.

Mello H., Panunzi A., Raso T. (eds), *Pragmatics and Prosody. Illocution, Modality, Attitude, Information Patterning and Speech Annotation* © 2011 Firenze University Press.

sample (Italian mini-corpus) was extracted from the C-ORAL-ROM Italian corpus (Cresti & Moneglia 2005; Cresti, Panunzi & Scarano 2005), and represent a larger part of the Italian informal corpus.

C-ORAL-BRASIL is a corpus of spontaneous speech of Brazilian Portuguese, coordinated by Tommaso Raso and Heliana Mello. The project is part of an international cooperation and constitutes the fifth branch of the European C-ORAL-ROM project (Cresti & Moneglia 2005). The architecture of the Brazilian corpus follows the same guidelines of the European corpora represented in C-ORAL-ROM, which ensures the comparability of both language resources. The informal section of C-ORAL-BRASIL comprises 139 texts and a total of 208,130 words in 21:08:00 of recording sessions, with a total of 34,167 terminated linguistic sequences (utterances). The informal portion is divided according to the context of the interactions: family/private (105 texts and 159,364 words) and public (34 texts and 48,766 words). Each of these sections is further equally subdivided according to the type of interaction: monologues, dialogues or conversations. Each subsection contains 1/3 of the texts. The diatopic variety represented in C-ORAL-BRASIL is the one of Minas Gerais state, in particular the metropolitan area of its capital Belo Horizonte[3].

The main goal of both the C-ORAL-ROM and the C-ORAL-BRASIL corpora is the documentation of the diaphasic variation, necessary to represent really spontaneous speech. Therefore, besides the variation between private/familiar and public contexts and among the three interactional typologies (monologues, dialogues and conversations), the corpora try to document the largest variation in terms of different interaction situations, so allowing a great variation of activity and, as a consequence, of different speech acts and information structures.

As in C-ORAL-ROM corpora, C-ORAL-BRASIL transcriptions incorporate the annotation of prosodic boundaries proposed by Moneglia & Cresti (1997). The annotation scheme segments the speech flow in two distinct levels. The first level deals with the demarcation of the fundamental entity in spontaneous spoken communication, that is the utterance. The utterance is signaled by a prosodic boundary that bears a conclusive value (terminal prosodic break) and conveys a speech act. The second level refers to the internal structure of the utterance, that can be built by one single tone unit (simple utterance) or by several tone units (compound utterance). Tone units within an utterance are prosodically signaled by boundaries with non-conclusive value (non-terminal prosodic break) (Moneglia & Cresti 1997; Moneglia & Cresti 2006).

[2] C-ORAL-BRASIL corpus will contemplate two major sections: one for informal speech and one for formal speech. The informal section is completed and the formal section is in compiling phase.

[3] More detailed information can be found at Raso & Mello (2010) and in press.

The Brazilian and European corpora have been designed to allow the study of illocutions and information structure of spontaneous speech. In order to allow the latter, the Brazilian mini-corpus received a tagging (complementary to the annotation of prosodic segmentation) that associates information functions to each one of the segmented prosodic units (Cresti 1987; Cresti 2000; Cresti & Moneglia 2010). The informational tagging is based in the model proposed by Language into Act Theory (Cresti 1987; Cresti 2000). This model was first implemented in the LABLITA corpus of Spontaneous Spoken Italian (Cresti 2006), from which the Italian corpus is derived.

The process and criteria of compiling the Brazilian mini-corpus is showed in section 2. In section 3 we present the methodology and tagset employed for the information structure annotation. Section 4 features some structural and informational characteristics of spontaneous spoken Brazilian Portuguese derived from the Brazilian mini-corpus. In section 5 we compare some of these results with the Italian mini-corpus. In section 6 the relationship between prosodic and informational annotation is discussed.

2. Strategy and criteria for compiling the Brazilian mini-corpus

In order to study the information structure, we need a corpus that identifies the informational functions of each prosodic unit; in other words, we must have an informationally tagged corpus. Unlike the tagging of part-of-speech, for which there are already many automatic tools, the tagging of information units is done manually. The information tagging of all C-ORAL-BRASIL texts, which comprises more than 61,000 information units, requires a considerable amount of time and human resources. For this reason, in a first stage, we selected a sample of the informal section of the C-ORAL-BRASIL corpus to receive informational tagging, thus enabling studies of informational nature.

The selection of texts followed criteria adopted to ensure a high quality database to perform information structure studies, but at the same time preserving the same basic structure of the entire corpus, so that the results obtained with the mini-corpus could be extrapolated to the whole corpus. Given the impossibility of balancing all the corpus variations in the mini-corpus, the parameters chosen as guidelines to achieve the best possible sample are the following (Raso & Mello 2009):

- Representativeness of typological branch. Dialogues and conversations should be 2/3 of the mini-corpus and monologues should be 1/3. The texts should be good exemplars of the context and text typologies: familiar/private and public dialogues, monologues and conversations.

- Highest possible range of communicative situations and activities. That means that speakers in different texts should perform different tasks, to ensure diaphasic variation.
- High acoustic quality. The quality is determined based on the absence (total or partial) of background noise, no feedback signal, voice clarity, good audio gain and low percentage of overlapping. The calculation of F0 curve must be (almost) always possible.
- Diversity of speakers. The goal is to have a balanced number of male and female voices and, if possible, also ages and school levels.
- Interesting text content. Texts with interesting content lead to higher attention of transcribers. Also, texts with interesting content increase the degree of informativeness within the sample.

The construction of the Brazilian mini-corpus involved the following steps.

1. Session recording, with the participants' consent, in digital format (wav).
2. Text transcription in CHAT format (MacWhinney 2000) with concomitant annotation of prosodic boundaries (Moneglia & Cresti 1997).
3. Review of the transcriptions, that includes the check for the appropriate application of the set transcription criteria (Mello & Raso 2009) and accurate annotation of prosodic breaks, always performed by a person other than the one who did the original transcription.
4. Text-to-spech alignment through software WinPitch (Martin 2005). Each audio file is aligned with the text according to the linguistic sequences marked by terminal prosodic boundary.
5. Informational tagging, performed on the aligned transcripts. During this phase, errors in the transcription and in the annotation of prosodic boundaries were also checked and corrected.
6. Two revisions of informational tagging and further correction of the transcripts.

The annotation of prosodic boundaries was validated in two occasions, once before the beginning of the transcription work and another when all transcriptions and the first revision were completed, but before the further revisions and the informational tagging. The final result of the validation reached a Kappa score agreement (Fleiss 1971) of 0,86, 0,87 for terminal breaks and of 0,78 for non terminal breaks (Raso & Mittmann 2009; Mello et al. in press).

Table 1 presents the information about each text of the Brazilian mini-corpus, indicating the text identification, the communicative situation, the number of male and female participants and duration of the audio file. The monologic group consists of narratives, descriptions and explanations. Monologues are highly elaborated texts, thus featuring less, but more complex, linguistic entities (utterances, illocutionary patterns and stanzas). Instead, conversations and dialogues comprise texts in which the speech is highly situated and entrenched in the immediate extra-linguistic context, and consequently they feature more utterances, with a less complex structure but with much more speech acts variation. As conversations are concerned, the first two represent the very common situation of friends just chatting.

Text names are composed by terms that indicate: language, context and text type. Thus we have ‘b’ for the Brazilian Portuguese, ‘fam’ to the family/private and ‘pub’ for public context, ‘cv’ for conversation, ‘dl’ for dialogue and ‘mn’ to monologue. Each text receives a double-digit sequential number that identifies it within the section to which it belongs.

Table 1. Situations recorded, number of male and female speakers and duration of texts

Text	Situation	M	F	Duration
Total		28	27	03:58:36
Conversations		15	9	01:07:28
bfamcv01	Chat between young friends	4	0	00:07:00
bfamcv02	Chat between elderly ladies	0	3	00:07:51
bfamcv03	Friends play snooker	5	0	00:06:50
bfamcv04	Friends play Pictionary	2	2	00:07:30
bpubcv01	Employees at a blood bank explain their work	1	3	00:08:30
bpubcv02	Political meeting	3	1	00:29:47
Dialogues		6	8	01:45:28
bfamdl01	Two friends do the groceries	0	2	00:14:39
bfamdl02	Two friends pack the recording equipment	1	1	00:07:26
bfamdl03	Couple takes a car trip	1	1	00:10:30
bfamdl04	Maids do the dishes	0	2	00:19:32
bfamdl05	Broker shows apartment to his sister*	1	1	00:11:28
bpubdl01	Engineer and construction worker at construction site	2	0	00:26:08
bpubdl02	Customer and salesman in a shoe store*	1	1	00:15:45
Monologues		7	10	01:05:40
bfammn01	Man tells an alleged true story about a snake	2	0	00:05:02
bfammn02	Grandmother tells grandson stories about her famous uncle	1	1	00:07:23
bfammn03	Father tells family two entertaining stories*	3	3	00:07:08
bfammn04	Woman tells about her experience in the hospital*	0	1	00:06:57
bfammn05	Woman shares the story about her daughter's adoption*	0	2	00:09:52
bfammn06	Man explains its professional trajectory	1	1	00:10:02
bpubmn01	Teacher evaluates her work at public school	0	2	00:19:16

* minor third party interventions.

The speakers' characteristics are almost perfectly balanced. In Table 1 are included also speakers that participate of the situation only for a few moments or that represent the interlocutors of the monologants. But if we consider the main speakers only, the balancing in term of uttered words is much better. The Brazilian mini-corpus features 23 speakers in conversation (one of them appears twice), 14 in dialogues and 7 in monologues. As far as gender is concerned, 25 are males and 19 are females, but the balancing in terms of words is almost perfect, since in conversations, where the number of words for each speaker in considerably smaller than in dialogues and specially in monologues, we have 16 males and only 7 females. The number of females is higher in dialogues (8 versus 6) and in monologues (4 versus 3). Age and school level are also balanced.

For age, we have in conversations 9 A speakers (from 18 to 25 years old), 9 B speakers (from 26 to 40 years old), 4 C speakers (from 41 to 60 years old) and 2 D speakers (more than 60 years old); in dialogues 4 A speakers, 3 B speakers, 6 C speakers and 1 D speaker; in monologues, 4 C speakers, 2 D speakers and 1 B speaker. For school level, speakers are divided in three different levels: level 1 refers to a school level up to incomplete primary school (no more than 7 school years); level 2 refers to a school level up to graduation, if the occupation of the speaker does not need the university degree; level 3 refers to a higher school level. In the mini-corpus, conversations feature 4 speakers with school level 1, 11 with school level 2 and 8 with school level 3; dialogues feature 2 speakers with school level 1, 7 with school level 2 and 5 with school level 3; monologues feature 3 speakers with school level 1, 2 with school level 2 and 2 with school level 3.

The most important feature of the Brazilian mini-corpus is its large diaphasic variation. As one can see in Table 1, the mini-corpus includes many different communicative situations. The diaphasic variation is an important parameter, on one hand because it is what ensures that the texts are really spontaneous and produced in natural contexts, and on the other hand, because diaphasic variation leads to variation in the information structure and in illocutionary values within the corpus.

The Brazilian mini-corpus maintains the same structure of the informal C-ORAL-BRASIL, divided into two sections, family/private and public situations, which are subdivided into conversations, dialogues and monologues. As in informal language perfect monologues are almost impossible, monologues are here defined as situations in which there is a clear predominance of textual elaboration by one of the speakers and almost no interaction. Dialogues are situations in which the linguistic exchange is focused on two informants (even if there are more minor intervenients) that produce a text highly entrenched in the extra-linguistic context. Conversations are much like dialogues, but they involve the active participation of three or more speakers. Table 2 shows the word distribution in each branch of the Brazilian mini-corpus.

Table 2. Number and proportion of words of the Brazilian mini-corpus

Context	Total		Conversations		Dialogues		Monologues	
Total	31318	100%	9774	31%	11331	36%	10213	33%
Family/private	23272	74%	6348	20%	8325	27%	8599	27%
Public	8046	26%	3426	11%	3006	10%	1614	5%

The Brazilian mini-corpus has a total of 31,318 words in 3:58:36 of recording. The distribution of words in each branch of the mini-corpus is showed in Table 2. In total, there is a balance regarding the percentage of words in each type of interaction: conversations have 31% of words, dialogues have 36% and monologues have 33% of total words.

It is important to say that, for many aspects, conversations and dialogues should be considered as one interactive typology versus monologues, that are a textual typology; therefore, a balanced mini-corpus should endure 2/3 of interactional typology and 1/3 of textual typology. The family/private context comprises 74% of the total number of words, and texts in public contexts represent only 26% of the total words in the mini-corpus. Due to the low representativeness of the public context, it is not possible to consider the context as a variable in studies based in the Brazilian mini-corpus.

3. Informational tagging

All 20 texts received informational tagging, using the set of informational units proposed by the Language into Act Theory and the Informational Patterning Hypothesis (Cresti 2000). In this framework, each utterance can be analyzed informationally. The only unit that is necessary and sufficient to build an utterance is the Comment unit, since it carries the illocutionary force of the speech act and gives prosodic and pragmatic autonomy to the utterance. The complex utterances consist of the comment unit and one or more units that accomplish different functions. These unit can be textual, when their function is to build the very text of the utterance, or dialogic, when their function is to support the interaction. The textual units, besides the Comment, are Topic (TOP), Appendix of Comment (APC), Appendix of Topic (APT), Parenthetical (PAR) and Locutive Introducer (INT). The dialogic units are Incipit (INP), Conative (CNT), Allocutive (ALL), Phatic (PHA), Expressive (EXP) and Discourse Connector (DCT).

Each unit is identifiable through three criteria: a functional criterion, a prosodic criterion and a distributional criterion; so, each unit has its specific function, its specific prosodic profile and its specific or preferential position in the utterance.

Other complex informational patterns are formed by Multiple Comments (CMM). In these cases two or more comments in the same utterance produce a rhetorical effect that causes that the two (or more) speech acts are interpreted as a whole. This is what happens in lists, comparisons, reinforcement, confirmation requests, among others types of multiple comments (Raso in press). Sometimes an informational unit can be segmented into more than one tone unit, characterizing the phenomenon of the Scanning unit (SCA). The scanning unit is due to difficulty in speech production, emphatic reasons or to articulatory necessity in case of too extended information units in terms of syllabic dimensions.

Finally, when there is less actional and interactional activity and the speakers builds a semantic text, the utterance is somehow dilated, giving rise to what is called *Stanza*. *Stanzas* are linguistic entities that do not correspond to the execution of one illocutionary force nor of a conventionalized rhetoric pattern, but to a broader linguistic activity, such as the construction of narratives and arguments. The *stanzas* are composed of sequences of Bound Comments (COB), whose junction is processual and not patterned. A complete listing of informational tags are shown in Figure 1. Later on this paper we will deep in the description of each information unit.

Textual information units		**Dialogic information units**	
COM	Comment	INP	Incipit
CMM	Multiple Comment	CNT	Conativ
COB	Bound Comment	PHA	Phatic
COB_s	Subordinator Comment	ALL	Allocutive
TOP	Topic	EXP	Expressive
TPL(n)	List of Topic: n indicates ordinal sequence	DCT	Discourse Connector
TOP_s	Subordinator Topic		
APC	Appendix of Comment	**Informationally empy units**	
APT	Appendix of Topic	SCA	Scanning
PAR	Parenthetic	EMP	Empty (incomplete units)
PRL	List of Parenthetic	TMT	Time Taking
INT	Locutive Introducer	UNC	Non identifiable
Further mark			
_r	Reported speech unit		

Figure 1. Tagset for the information units

Before they start the tagging, the annotators went through a phase of training, exercises and discussions that involved the project coordinator and the researchers of the LABLITA lab. The goal was not only to enable annotators with respect to the theoretical tools, but also to establish a standard of uniformity and consistency. The annotators also went through a statistical evaluation of the degree of agreement before beginning the informational tagging task. All annotators independently tagged a dialogue with 120 utterances (171 tone/information units) and a monologue with 70 utterances (372 tone/information units).

The overall results of the inter-rater agreement test (Kappa Statistics) were 0.62 for the utterance and 0.73 for tone/information unit. A more detailed analysis pointed out that the disagreement cases were restricted to just a few information tags. These problems could be managed in the revision phase. Most of the problems encountered in the tagging and showed by the Kappa Statistics test involved the two specific information units: Multiple Comments and Bound Comments, which are the less studied units so far.

Tagging went through two distinct phases of review. The first was conducted by one annotator, always different from who had originally tagged the text, together with the coordinator of the project. Sometime later, the informational tagging was again reviewed by the project coordinator in conjunction with a member of the European project (C-ORAL-ROM), which is the most experienced person in relation to informational tagging based on the Language into Act Theory[4]. This last revision had both the goal to better the accuracy of the informational tagging and to ensure consistency with the tagging of the Italian corpus.

4. Structural and informational features

The first measurements to be observed in order to obtain a better knowledge of spontaneous speech are the distribution of dialogic turns, the number of utterances and the number of tone/information units in the sample and its branches. The averages of utterances per turns and of tone/information unit per utterance allow to evaluate the degree of interaction of the texts. The lower these numbers, the higher the interaction degree.

The average of utterances per turn is a measurement that reflects the alternation of the turns during the interaction: therefore, if the turns are short in terms of utterances, this means that the interactivity is high; when the turns show many and longer utterances, this reflects a lower degree of interactivity. As far as the average of tone/information unit per utterance is concerned, we can observe that the higher the number of tone unit per utterance is, the more complex the utterances are; a high number of very complex utterances is typical of interactions with a low degree of interactivity. The reason is that the utterance complexity goes together with the amount of textual information units; and the more text we have in the interaction, the less percentage of illocution, i.e. actionality, and therefore interactivity, we have. Conversations and dialogues show turns with a lower number of utterances and utterances with a lower number of tone/information units.

Table 3 shows these values for each text and for each interactional typology. The Table also presents other values: the average values of utterances per turns

[4] We thank Ida Tucci of the LABLITA lab for her collaboration.

calculated considering only the concluded utterances, and the average value of tone units per utterances calculated considering only the concluded tone units.

Looking at the data of Table 3, we can observe several characteristics of the texts and their structure. First of all, it is evident the difference between dialogues and conversation on one hand, and monologues on the other hand, in terms of number of turns average. While conversation and dialogues have almost the same number of turns (respectively 1333 and 1371), monologues show a much lower number of turn (250).

The same opposition between dialogues and conversations on one hand and monologues on the other hand can be confirmed with respect to other measurements, as already observed for the C-ORAL-ROM languages by Cresti (2005):

- The average of concluded utterances per turn is similar between conversations and dialogues (respectively 1,39 and 1,66), while it is much higher for monologues (3,68);
- The average of tone units per utterance also is similar for conversations and dialogues (1,71 and 1,54) and much higher for monologues (2,94);
- The number of retracting phaenomena is also similar for conversations and dialogues (253 and 228) and much higher for monologues (388).

Table 3. Structural features of Brazilian mini-corpus

Text typology	**Dialogic turns (DT)**	**Interrupted sequences**	**Concluded sequences (CS)**	**CS/DT**	**Retracted units**	**Informative tone units**	**IU/CS**
Total	2954	441	5043	1,71	869	9384	1,86
Conversations	1333	191	1848	1,39	253	3164	1,71
bfamcv01	159	41	207	1,3	46	441	2,13
bfamcv02	239	29	356	1,49	36	579	1,63
bfamcv03	185	10	296	1,6	38	467	1,58
bfamcv04	323	43	422	1,31	28	645	1,53
bpubcv01	265	32	323	1,22	35	611	1,89
bpubcv02	162	36	244	1,51	70	421	1,73
Dialogues	1371	176	2275	1,66	228	3513	1,54
bfamdl01	338	24	542	1,6	19	781	1,44
bfamdl02	176	35	247	1,4	56	453	1,83
bfamdl03	172	38	300	1,74	41	505	1,68
bfamdl04	123	9	244	1,98	11	367	1,5
bfamdl05	239	40	391	1,64	46	566	1,45
bpubdl01	158	14	262	1,66	32	407	1,55
bpubdl02	165	16	289	1,75	23	434	1,5
Monologues	250	74	920	3,68	388	2707	2,94
bfammn01	19	8	98	5,16	70	245	2,5
bfammn02	95	13	171	1,8	57	477	2,79
bfammn03	48	9	135	2,81	59	353	2,61
bfammn04	26	8	181	6,96	21	446	2,46
bfammn05	31	18	135	4,35	56	401	2,97
bfammn06	6	4	72	12	47	328	4,56
bpubmn01	25	14	128	5,12	78	457	3,57

These measurement allow us to establish a first opposition between dialogic texts (conversations + dialogues) and monologic texts. This opposition will be confirmed analyzing the information structure of these two major typologies. Nevertheless this does not eliminate completely the differences between conversations and dialogues.

First of all it is necessary to note that in number of words the two typologies are not perfectly balanced (since the most important balancing is due to the opposition between dialogic and monologic typologies): in the mini-corpus, we have 9843 word for conversations and 11371 words for dialogues (since we have 7 dialogues and 6 conversations). This does not reflect any significant difference in term of turn dimensions, as conversations have 7,38 words per turn and dialogues have 7,18 words per turn. But if we observe the number of interrupted sequences, we note that its rate (number of interrupted sequences divided for the number of words) is 1,94 in conversations and only 1,54 in dialogues. Similarly, the rate of retractings is 2,57 in conversations and only 2,0 in dialogues. This means that the higher competition for the turn in conversation causes a higher number of fragmentation phaenomena.

Another interesting difference is the higher rate of tone units per turn in conversations (1,71) with respect to dialogues (1,54). This difference seems to reflect the fact that in conversations it is easier to find parts in which one speaker articulates more complex utterances, but we have also to consider that in the mini-corpus we have 2 conversations without a specific activity performed by the speaker, which can also contribute to a less actional interaction.

In fact, different text typologies, specially the opposition between dialogic typologies and monologic typology, have important consequences on the information structure of spoken discourse. Table 4 shows some important values in order to distinguish the structure of conversations, dialogues and monologues.

The data presented in this Table was extracted through the search interface of DB-IPIC, a database in XML format implemented by Panunzi e Gregori (2011; also in this volume). It allows the study of information units in spoken corpora annotated according to the Information Patterning Theory (Cresti 2000; Moneglia & Cresti 2006; Scarano 2009).

We can observe that the highest level of difference is the percentage of presence of the three units of reference: utterance, illocutionary pattern and stanza. As the data show, clearly more than 80% of conversation and dialogue structure is built by utterances, 10% by illocutionary patterns and only a very little part by stanzas, which, moreover, are usually very simple, in term of structure. The differences between conversations and dialogues are very little, but we will come back to this later.

The most important aspect now is to note how different is the composition of the monologic typology. It features only 66% of utterances, 8% of illocutionary patterns but 25% of stanzas, which are often very complex. So we can say that stanza is a reference unit typical of monologic texts, and this is a very important feature of the informational complexity of this typology.

Table 4. Informational features of the Brazilian mini-corpus

Informational typologies	**Conversations**		**Dialogues**		**Monologues**	
Total linguistic entities	1855	100,0%	2304	100,0%	950	100,0%
Total utterances	1534	82,7%	1972	85,6%	633	66,6%
Simple utterances *COM*	1095	71,4%	1452	73,6%	351	55,5%
Simple scanning utterances *COM + SCA, TMT, EMP*	91	5,9%	121	6,1%	63	10,0%
Compound utterances with dialogic units *COM + ALL, CNT, DCT, EXP, INP, PHA*	196	12,8%	232	11,8%	63	10,0%
Compound utterances with textual units *COM + APC, INT, TOP, TPL, APT, PAR, PRL*	108	7,0%	125	6,3%	100	15,8%
Mixed compound utterances *COM + textual and dialogic units*	44	4,0%	42	2,9%	56	16,0%
Total illocutionary patterns	202	10,9%	225	9,8%	77	8,1%
Simple illocutionary patterns *2 or more CMM*	147	72,8%	148	65,8%	34	44,2%
Simple scanning illocutionary patterns *2 or more CMM + SCA, TMT, EMP*	13	6,4%	19	8,4%	10	13,0%
Compound illoc. patterns with dialogic units *2 or more CMM + ALL, CNT, DCT, EXP, INP, PHA*	24	11,9%	30	13,3%	8	10,4%
Compound illoc. patterns with textual units *CMM + APC, INT, TOP, TPL, APT, PAR, PRL*	14	6,9%	20	8,9%	21	27,3%
Mixed compound illocutionary patterns *2 or more CMM + textual and dialogic units*	4	2,0%	8	3,6%	4	5,2%
Stanzas *at least one COB + COM*	119	6,4%	107	4,6%	240	25,3%

But the complexity of the monologic typology is also testified by analizing the internal structure of the utterance. In conversations and dialogues, the most part of the utterances are simple utterances, while in monologues the proportion of compound utterances is much higher.

It is interesting also to observe the percentage of simple scanning utterances, that means utterances built by the comment and one or more informationally empty units, like scanning units or time taking or not concluded units. It is important also to observe that the scanning simple utterances have a much higher weight in monologues than in the dialogic typology. This depends on at least two factors: first, the higher fragmentation phaenomena in monologues, due to the processual construction of a more complex text, and second, to the fact that many of these cases, certainly more than in dialogic typology, are due to the interruption of a compound unit, and therefore are included inside simple utterances only because the interruption happens before the realization of a full informational unit.

The monologic informational complexity can be confirmed by another important aspect: the relevance of textual units in building compound utterances. In Table 4 there is a differentiation among compound utterances with dialogic units, compound utterances with textual units and mixed compound utterances. This last category includes all the compound utterances that have both textual and dialogic units inside. For our purpose here, utterances that have textual units, independently if they have also dialogic units or not, will be considered as one unified category and compared with the compound utterances with only dialogic unit (besides, of course, the comment unit). Compound utterances with only dialogic units are more frequent in conversations and dialogues, where they sum respectively 12,8% and 11,8% of all the utterances, that are respectively 82,7% and 85,6% of all the reference units. Only 11,% of the utterances in conversations and 9,2% in dialogues are compound utterances with at least one textual unit. In monologues, what happens is much different: just 10% of the compound utterances are build only by dialogic units, while 31,8% have at least one textual unit.

If we now analyze the illocutionary patterns, we realize that they are more frequent in dialogic typologies, but also that they are more complex in textual typologies. In fact, without considering the simple scanning illocutionary patterns (that may depend on different reasons), we can observe that only 9,% of the illocutionary patterns in conversation and 12,5% in dialogues have textual units, while in monologues illocutionary patterns with textual units reach 32%.

All these measurements allow us to conclude that dialogic typologies are basically built on a sequence of simple utterances or illocutionary patterns. This means that these typologies are strongly based on alternation of the illocutionary force. The high presence of dialogic units shows that if the speaker needs more units than the illocutionary ones, they are still directed to the interlocutor in order to guaranty the interaction (dialogic units), and do not build the text of the utterance. The presence of textual units is in fact very low. On the contrary all the measurements in Table 4 lead us to conclude that monologic typology has a completely different structure. The very high weight of stanzas and of compound utterances with textual units shows that the importance of the illocutionary force is much lower, while the importance of really informative units, that means units that build the text of the reference unit, is very high.

This can be explained with the fact that the basic activity the speakers perform when interacting and when build a monologic text is different: the interaction is an alternation of actions that the speakers do toward their interlocutor, and for this they need, besides the illocutionary force, also dialogic unit that provide the regulation of the channel and that of the social cohesion between the speakers. On the other hand, monologues are principally elaborated texts (argumentations, narratives, explanations, descriptions) built by only one speaker. He may have a certain degree of interaction with the listener(s), but his activity is mainly that of organizing and giving voice to his thinking, not to perform actions pulsioned during the interaction.

While the dialogic texts develop on the basis of interaction, monologues are a process of text construction by just one speaker. In dialogic texts the speaker does not have a mental project to develop, and interacts with the interlocutor depending on unforseeable interlocutor's action. In monologues the speaker does have a mental project, for instance to tell a story or to explain something, and this lead to a complex mental process in which the illocutionary force weakens and the semantic text construction takes, to a certain extent, its place.

5. Information structure in Brazilian Portuguese and in Italian

5.1 The characteristics of the information units

Before making a very general comparison between the Brazilian and Italian mini-corpora, it is necessary to offer some more informations about the function, the prosodic profile and the distribution of the information units[5].

The textual units build the text of the utterance. The only unit that is necessary and sufficient to build an utterance, as it carries the illocutionary force, is the Comment unit (COM). When this unit is patternized with another illocutionary unit gives rise to the Multiple Comment (CMM). In prosodic terms, they are root units ('t Hart et al. 1990), and are the only unit that has prosodic and pragmatic autonomy. Its prosodic profile changes according to the illocution that is conveyed (Firenzuoli 2003; Moneglia 2011) and always bears a functional nucleus, that is the prosodic portion that conveys the specific illocutionary value (see Mello & Raso in this volume; Cresti in this volume). Its distribution is free. Also the Bound Comments COB) are root units, but with a weakened illocutionary value. They appears in

[5] More detailed informations about the different information units can be accessed in Cresti (2000), Raso (in press) and in the bibliography about the specific unit.

stanzas (Cresti 2009) and are typical of monologic text, where *stanzas* can be very big and complex, organized in subpattern around each Bound Comment. The Bound Comment ends with a continuity prosodic signal, that marks that the reference unit is not concluded and that the illocutionary force must be interpreted inside a broader reference unit.

Figure 2 shows the distribution of the different root units in the three text typologies in the Brazilian mini-corpus. Once again we notice that the different unit have a similar distribution in conversations and dialogues, but a very different one in monologues.

While the greatest part of root units for conversations and specially dialogues is the Comment unit, for monologues Bound Comments have a very important role, reaching almost 1/3 of all the root units.

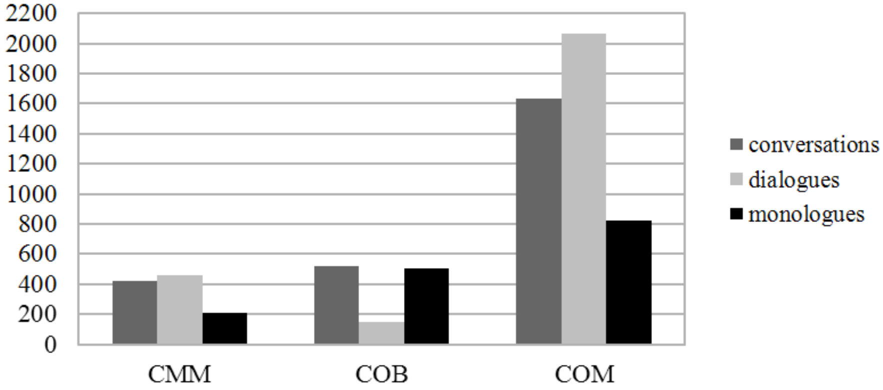

Figure 2. Distribution of the root units in the three text typologies

Actually, the real weight of Bound Comments in monologues is much bigger than Figure 2 shows. In fact, it is very common that the interlocutor constantly signals his attention by uttering simple utterances like hum hum // or exclamations that show his participations in the interaction. All these cases, which should not be considered within the monologue structure, are computed in the graphic as Comment unit. On the contrary, the weight of Bound Comments in conversations is only 20% and in dialogues 5,5%. As far as Multiple Comments are concerned, they concentrate in the dialogic typologies. Comparing conversations and dialogues, it is possible to observe that conversations have a little less Comment units and a higher presence of bound comments.

A compound information pattern contains one (or more) root units and normally has also textual or dialogical information units.

The textual information units are:

- The *Topic* (TOP) unit (Signorini 2003) is the most important unit of an information pattern. Its function is to define the cognitive perspective, that means the semantic dominion, of the illocutionary force. Prosodically, it is the only unit, besides the comment, that bears a functional nucleus, despite the fact that, like all information units except the comment, it is not pragmatically interpreTable in isolation. The nucleus is always, entirely or partially[6], positioned on the right of the unit (Firenzuoli & Signorini 2003; Raso et al. forthcoming). Its distribution is always on the left of the comment.
- The *Appendix* unit integrates the text of the Comment (Appendix of Comment – APC) or of the Topic (APT). Prosodically the Appendix has a descendent or flat profile. The APT can show movement, but without any focus. Their distribution is always on the right of the Comment or of the Topic (Raso & Ulisses 2008; Ulisses 2008; Tucci 2006).
- The *Parenthetic* (PAR) has the metalinguistic function to make a commentary about the utterance or part of it. Its profile is flat, with a lower (or rarely higher) F0 level with respect of the rest of the utterance, and a higher speech rate. It can occupy any position, even inside another textual unit, except the beginning of the utterance (Tucci 2004; Tucci 2009).
- The *Locutive Introducer* (INT) has the function to introduce a list of topics and specially an illocutionary pattern with a meta-illocutionary value, outside of the deictic coordinates of the utterance (Corsi 2009; Maia Rocha 2010; Maia Rocha & Raso 2011). One very important function of the INT is therefore that of marking the suspension of the pragmatic coordinates of the utterance introducing a different *hic et nunc*. Prosodically, INTs have a descendent profile, with a much lower F0 frequency with respect to the meta-illocution that follows, producing a clear F0 contrast that marks also prosodically the suspension of the pragmatic coordinates, and with a much higher speech rate. Its distribution is before the introduced units.

Figure 3 shows the distribution of the different textual units in the three text typologies of the Brazilian mini-corpus.

[6] Some Topic prosodic profiles have two semi-nuclea. In this case, the preparation (that depends on the syllabic dimensiono f the locutive contet) can be positioned between the two nuclear portions.

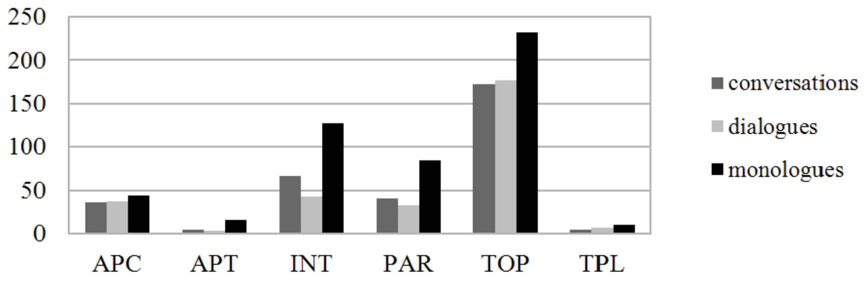

Figure 3. Distribution of the textual units in the three text typologies

It is noticeable that all textual units have a much higher presence in monologues. This is specially true for Topics, that are much more necessary in situations that cannot have the pragmatic situational context as an immediate reference for the illocutions, like in narratives, descriptions or argumentations, for the Locutive Introducer, since in monologues it is much higher the use of meta-illocutions, and for parenthetical, that allows the speaker to modalize and to make commentary on the textual content of the utterance.

Again, it is possible to notice a small difference between conversation and dialogues, always with conversations showing, in a very little proportion, the tendency to present some characteristics of monologues.

The dialogic units (Frosali 2008) are very different, with many respects, from the textual ones. Their function is not that to build the text of the utterance, but that of controlling the interaction. The dialogic units are:

- The *Incipit* (INP) has the function of beginning the turn or the utterance with contrast with the previous one; its prosodic profile is ascendent-descendent (or only ascendent or only descendent) reaching a high F0 value with a very short duration and high intensity; it opens the utterance.
- The *Phatic* (PHA) has the function to signal that the channel is open, with a very short and flat or descendent profile, and with low intensity; its position is free.
- The *Allocutive* (ALL) has two functions: to individualize the interlocutor, but specially to mark the social cohesion with him; its prosodic profile is descendent or slightly modulated, with standard duration and intensity; it must not be confused with the recall illocution (Raso & Leite 2010).
- The *Expressive* (EXP) has the function to support emotionally the illocution; its profile may vary, but it is usually modulated, with standard duration and intensity.

- The *Conative* (CNT) has the function to press the interlocutor to do or quit doing something; its profile is descendent, with short duration and high intensity.
- The *Discourse Connector* (DCT) has the function to open the utterance without contrast with the previous one, or to connect the subpatterns inside a *stanza*; its profile is flat or modulated, with high intensity and long duration.

Figure 4 shows the distribution of dialogic units in the three text typologies of the C-ORAL-BRASIL mini-corpus. The distribution of the dialogic unit is very interesting to show some specific aspects of the three text typologies. First, we can observe that is frequent an opposition between dialogic typologies and monologues. This is clear with respect to Conatives, Expressives and Discourse Connectors. In the first two cases, the dialogic typologies show a clearly higher presence of these units, but for discourse connectors the opposite happens. We will be back on this later.

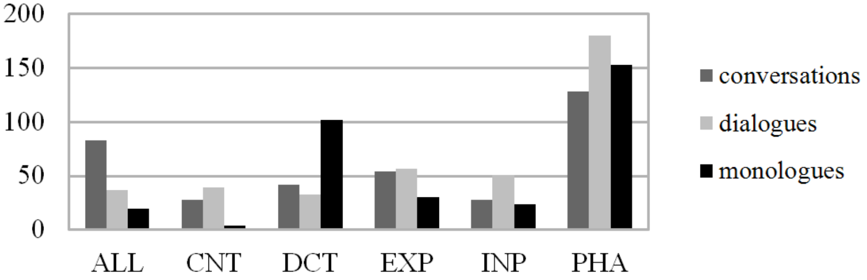

Figure 4. Distribution of the dialogic units in the three text typologies

Concerning the Allocutives, there is a sort of scale that goes from the highest presence in conversations to the lowest presence in monologues. This distribution of this unit depends on some well-studied factors. Allocutives are a strongly dialogic unit, as they are used always to support the interaction. They have, as already said, two main functions: that of individualizing the interlocutor and of marking the social cohesion with him. This last function is equally strong in conversations and dialogues, but very low in monologues. The first function does not make sense in dialogue but only in conversation. So this explain the fact that conversation has a higher use of allocutives with respect to dialogue. But what is the function of allocutives in monologues? Monologues, specially narratives, have a high amount of reported speech; in reported speech allocutives are used to indirectly signal to the interlocutor who are the reported speaker and the reported interlocutor.

The distribution of incipit and phatic still needs to be studied.

The distribution of Discourse Connector reflects the specific function of this unit, for many respects different from the other dialogic unit. As we said, its function is to mark continuity between two utterances, but also to connect subpatterns in a *stanza*. As the former function is common to the three text typologies, the last one is typical of the monologic typology, where *stanzas* are much more present and much more complex.

5.2 A first comparison with the Italian tagged mini-corpus

The informational tagging of the Italian C-ORAL-ROM corpus began much earlier than the tagging of C-ORAL-BRASIL. Therefore, in order to study the information structure in a cross-linguistic perspective, part of the Italian tagged corpus was extracted to be compared with the 20 tagged texts of the C-ORAL-BRASIL corpus. As the Brazilian mini-corpus is highly actional, to turn the Italian mini-corpus comparable to the Brazilian one, the priority was to maintain the same proportion between dalogic and monologic typologies and to maximize the actionality of the text, meaning with this, the maximum number of varieties of activities performed by the speaker while interacting. The composition of the Italian mini-corpus is that presented in Table 5.

The Italian mini-corpus is a little bigger, in terms of words, than the comparable Brazilian one, but its balancing, with respect to the two priorities (1/3 of monologic and 2/3 of dialogic texts, and maximization of different actional texts) is almost perfect. Since the Italian mini-corpus was adapted to the Brazilian mini-corpus, it cannot maintain the almost perfect balance with respect to the speakers' characteristics.

Here, we will only propose some general observations comparing the two mini-corpora. A better and deeper comparison needs a specific dedicated study. Table 6 shows for Italian the same data that Table 4 shows for Brazilian.

Table 5. The Italian mini-corpus

Text	**Situation**	**M**	**F**	**Words**
Total		23	31	34208
Conversations		9	11	10141
ifamcv01	relatives talk while browsing through family photos	1	2	
ifamcv09	friends explain the game Mastermind	3	0	
ifamcv15	family talks with child during lunch preparation	2	3	
ipubcv01	exchanging ideas during a meeting of a voluntary association	1	4	
ipubcv05	chat in a ironmonger while shopping	2	2	
Dialogues		5	13	12435
ifamdl04	interview of an artisan in his leather workshop	1	2	
ifamdl12	friends at home making a cake	0	2	
ifamdl15	beautician and customer in the beauty-center	0	2	
ifamdl17	two friends develop photos in a dark-room	1	1	
ifamdl19	father gives driving lesson to his daughter	1	2	
ifammn17*	professional explanation to a colleague about office-work	0	2	
ipubdl02	proposal of an insurance policy	0	2	
ipubdl05	teachers' meeting at the school office	2	0	
Monologues		9	7	11632
ifammn02	interview with an old partisan at his home	2	0	
ifammn05	elderly woman tells life story to her relatives	1	2	
ifammn08	narrative to a relative about the honeymoon	0	1	
ifammn03	an after-dinner travel tale to friends	2	2	
ifammn14	interview with a retired travelling-salesman	1	1	
ipubmn01	political speech at a political-party meeting	2	0	
ipubmn04	interview with an employee of the Poggibonsi municipality	1	1	

*Labeled as monologue but is acctually a dialogue

Table 6. Information features of the Italian mini-corpus

Informational typologies	**Conversations**		**Dialogues**		**Monologues**	
Total linguistic entities	1769	100,0%	2054	100,0%	1195	100,0%
Total utterances	1481	83,7%	1714	83,4%	842	70,5%
Simple utterances *COM*	987	66,6%	1169	68,2%	329	39,1%
Simple scanning utterances *COM + SCA, TMT, EMP*	95	6,4%	126	7,4%	90	10,7%
Compound utterances with dialogic units *COM + ALL, CNT, DCT, EXP, INP, PHA*	144	9,7%	178	10,4%	116	13,8%
Compound utterances with textual units *COM + APC, INT, TOP, TPL, APT, PAR, PRL*	172	11,6%	168	9,8%	186	22,1%
Mixed compound utterances *COM + textual and dialogic units*	83	8,4%	73	6,2%	121	36,8%
Total illocutionary patterns	183	10,3%	172	8,4%	80	6,7%
Simple illocutionary patterns *2 or more CMM*	106	57,9%	93	54,1%	25	31,3%
Simple scanning illocutionary patterns *2 or more CMM + SCA, TMT, EMP*	23	12,6%	17	9,9%	10	12,5%
Compound illoc. patterns with dialogic units *2 or more CMM + ALL, CNT, DCT, EXP, INP, PHA*	15	8,2%	22	12,8%	14	17,5%
Compound illoc. patterns with textual units *CMM + APC, INT, TOP, TPL, APT, PAR, PRL*	31	16,9%	28	16,3%	21	26,3%
Mixed compound illocutionary patterns *2 or more CMM + textual and dialogic units*	8	4,4%	12	7,0%	10	12,5%
Stanzas *at least one COB + COM*	105	5,9%	168	8,2%	273	22,8%

We can confirm that for Italian, dialogic texts behave in similar way, while monologic texts present very different measures. We can observe that the proportion of utterance is the same comparing dialogues and conversations. This allows us to hypothesize that the small differences found between these two typologies in the Brazilian mini-corpus are due to the presence of two conversations in which the speakers do not perform any specific activity, pushing therefore some measurements in the direction of monologic values. We can also observe that monologues in the Italian mini-corpus present a little less stanzas and more utterances, but also a little less illocutionary patterns. In any case, these can be considered not to be significant differences.

A more significant difference is the fact that in Italian the percentage of simple utterances is much lower than in Brazilian. While Brazilian shows 71.4% of simple utterance in conversation, 73.6% in dialogue and 55.5% monologue, in Italian these measurements are, respectively, 66.6%, 68.2% and 39.1%, what seems to lead to a more complex informational structure for this language. This hypothesis is strengthened by the fact that the number of textual compound utterances is also higher in Italian. While Brazilian shows a percentage of 11.00%, 9.2% and 31.8% of textual compound utterances respectively for conversations, dialogues and monologues, Italian presents 20.0%, 16.00% and 58.9%. The same happens for illocutionary patterns: compound illocutionary patterns are much more common in Italian, while simple illocutionary patterns are much more common in Brazilian. Differences in terms of stanzas do not seem significant. Figure 5 shows the proportion of root units in the Italian mini-corpus.

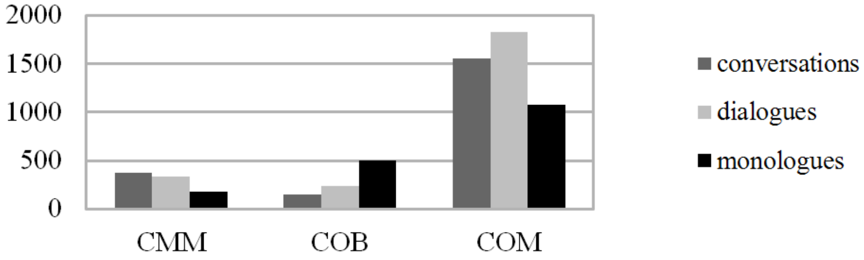

Figure 5. Distribution of the root units in the three text typologies in Italian

With respect to Brazilian root units and its distribution in the different branches of the mini-corpus, it is noticeable a lower number of illocutionary patterns: 10.3%, 8.4% and 6.7% respectively in conversations, dialogues and monologues, versus 10.9%, 9.8% and 8.1% in Brazilian. On the contrary, the number of bound comments is much higher (with the exception of conversations).

Figure 6 shows the distribution of textual units in the Italian mini-corpus and corresponds to Figure 3 for Brazilian. We can observe the much higher number of all textual units in Italian, with the only exception of the Locutive Introducers.

The fact that Locutive Introducers are in contratendential distribution with respect of the other textual units is something that must be explained: first of all we can observe that in Italian the INTs distribution does not vary much in the three typologies, even if monologues have more INTs and dialogues have less INTs; in the Brazilian mini-corpus the number of INTs in monologues is much higher than in the other typologies.

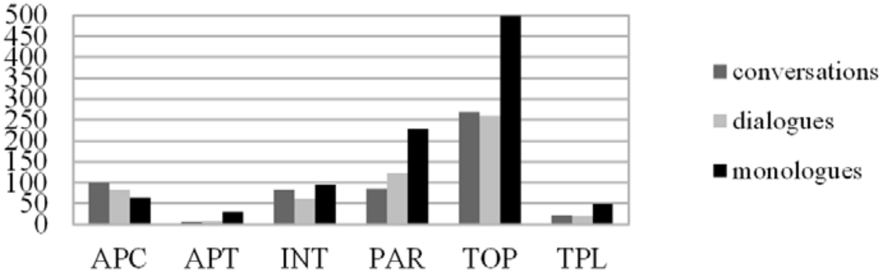

Figure 6. Distribution of the textual units in the three text typologies in Italian

A hypothesis that should be tested is that reported meta-illocutions, and specially reported speech, are much more frequent in Brazilian, since they represent a more pragmatic and less textual strategy of text building. Another interesting difference between the two mini-corpora with respect to textual units is the inverted distribution in the different typologies of the APCs. While Brazilian has more APC in monologues and less in conversations, Italian shows more APCs in conversations and less in monologues.

Figure 7 shows the distribution of the dialogic units in the Italian mini-corpus, and corresponds to Figure 4 for Brazilian.

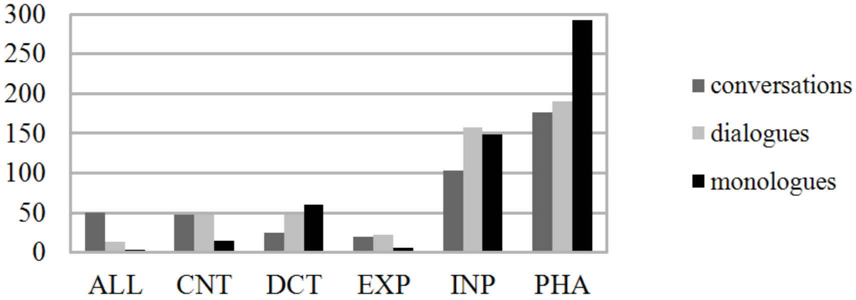

Figure 7. Distribution of dialogic units in the three text typologies in Italian

The different distribution of dialogic units in the two mini-corpora allows for many considerations. First of all it is important to emphasize the cultural relevance of dialogic units. They have the function to govern the interaction, and this is a very sensible to cultural characteristics function.

A study about allocutive in Italian, Spanish, European Portuguese and Brazilian Portugues (Raso & Leite 2010) shows that Brazilian Portuguese and European

Portuguese have a very different way to use this unit, with a difference between them higher than the difference that they show with respect to Spanish and Italian. Comparing Brazilian and Italian with respect to all the dialogic units, we note that Brazilian uses much more allocutives and expressives, while Italian uses much more conatives and incipits. The very high presence of phatics in Italian monologues is another remarkable difference. The last difference is the very high number of DCTs in monologues. These differences should still be better studied.

6. Annotation of prosodic boundaries and informational tagging

In this section we discuss the relationship between the annotation of prosodic boundaries and the identification of the informational value for the prosodic units in the Brazilian mini-corpus[7]. This research is necessary to the extent that the Brazilian mini-corpus transcripts had not undergone a revision after the text-to-speech alignment as the rest of the C-ORAL-BRASIL. Thus, during the informational tagging, annotators add or remove either words or prosodic breaks. In several cases, they also change the value of a prosodic break (for instance, from terminal to non-terminal or *vice versa*). Thus, this analysis aims to assess to what extent these changes were made in the prosodic annotation during tagging, and also discuss the change in the annotation with relation to specific information functions.

For this analysis we used two versions of the Brazilian mini-corpus. The first version consists of the transcripts after they passed through a first revision. The second one is the final informationally tagged version of the Brazilian mini-corpus. The total of analyzed transcripts amounts 40 texts. Each text went through an automatic processing through R computational tool (R Development Core Team 2010) in order to be prepared for data mining and statistical analysis. In a spreadsheet, each first version transcript was aligned, word by word, with the corresponding second version transcript. Naturally the versions of each text presented a different word numbers, due to word inclusions and exclusions in the final version. As the inclusion or exclusion of words can alter the annotation of prosodic breaks, changes in transcripts at the segmental level were also controlled.

After alignment, the sample adds up to a total of 31,750 tokens. Each token corresponds to a word boundary, considering the words of both versions. Of this total, 11,200 (35%) positions had a prosodic break either in the first or in the second version. Considering only these positions, we noticed that 6% of the tokens (651 cases) are involved in some sort of alteration in the segmental level, like additions, deletions and corrections of words (see Table 7).

[7] For a detailed description of the methodology for segmentation e its validation in the C-ORAL-BRASIL corpus, see Raso & Mittmann (2009), Mello et al. (in press).

Table 7. Types and frequencies of positions with annotation of prosodic breaks

Position type	**Freq.**	**%**
Total positions with prosodic breaks	11200	100%
Positions with segmental changes	651	6%
Word corrections	365	3%
Word inclusions	213	2%
Word exclusions	73	1%
Positions without segmental changes	10549	94%
Without changes in prosodic breaks	9175	82%
With changes in prosodic breaks	1374	12%

We do not consider for the analysis the positions in which there was any kind of modification at the segmental level. In this way, we eliminate possible changes in the annotation of prosodic breaks due to additions or deletions of words. Thus, the total analyzed data equals 1,374 tokens. Those correspond to the instances in which, at the same time, there were no segmental changes but that presented changes on the annotation of prosodic breaks.

As shown in Table 7, during the informational labeling, annotators made changes in 12% of the prosodic breaks. This value is high, nevertheless we must take into account that the transcripts underwent only one phase of revision before informational tagging, while the rest of the C-ORAL-BRASIL informal corpus passed by at least 4 revisions.

Changes include the addition and deletion of prosodic breaks, as well as the modification of the prosodic breaks value. The changes made during the informational tagging are summarized in Table 8.

Considering break exclusions (26% of total changes), one can notice that an irrelevant percentage of those relates to terminal breaks (0.29%) and to retracting and interruption (both equals 0.95%). Almost all the exclusions consist of non-terminal breaks deletions (24.09% and 331 cases).

Table 8. Types and frequencies of prosodic break changes

Type of prosodic break change	**Freq.**	**%**
Total positions with changes in prosodic breaks	**1374**	**100.00%**
Exclusion of prosodic break	**361**	**26.27%**
Terminal	4	0.29%
Non-terminal	331	24.09%
Interruption	13	0.95%
Retracting	13	0.95%
Inclusion of prosodic break	**375**	**27.29%**
Terminal	11	0.80%
Non-terminal	355	25.84%
Interruption	4	0.29%
Retracting	5	0.36%
Modification of prosodic break type	**638**	**46.43%**
Terminal → non-terminal	354	25.76%
Terminal → interruption	32	2.33%
Terminal → retracting	2	0.15%
Non terminal → terminal	90	6.55%
Non terminal → interruption	19	1.38%
Non terminal → retracting	24	1.75%
Interruption → terminal	27	1.97%
Interruption → non terminal	14	1.02%
Interruption → retracting	45	3.28%
Retracting → terminal	5	0.00%
Retracting → non terminal	22	1.60%
Retracting → interruption	4	0.29%

Most non-terminal break deletions (around 57%) are due to the inappropriate association of prosodic boundaries and discourse markers. Examples (1) and (2) below illustrate such occurrences.

(1) então / vamo passar lá // (bfamdl05) first version
então vamo passar lá // (bfamdl05) final version
[so let's go there]

(2) mas é isso aí / o' // (bfammn01) first version
mas é isso aí o' // (bfammn01) final version
[so this is it see]

What happens is that many lexical items, especially in initial position in the utterance, are candidates to be discourse markers, like 'então' (*so*), 'aí' (*so*), 'mas' (*but*), ‘e’ (*and*) and several other items. These are all items with low phonetic consintency, that may be realized very quickly; after them it is possible, but not necessary, that a prosodic break is realized, giving to them the status of discourse markers. As they are not syntactically compositional with the rest of the utterance, it is very likely that a boundary is perceived and attributed to prosodic aspects even when there is not any prosodic reason for this. This represents the typical case in which revisions reduce a wrong annotation.

Other significant exclusions (14%) are related to the false association between prosodic units and syntactic units. Non-terminal breaks were removed from the final version in contexts where a syntactic limit, such as clause ending, was falsely interpreted as containing also a prosodic boundary. See examples (3) and (4) below.

(3) eu ditando / e o Tommaso escrevendo // (bfamdl01) first version
eu ditando e o Tommaso escrevendo // (bfamdl01) final version
[I dictating and Tommaso writing]

(4) essa é a rua / que nós vimo // (bfamdl05) first version
essa é a rua / que nós vimo // (bfamdl05) final version
[this is the street that we saw]

The results show that almost all changes were related to non-terminal prosodic breaks. This is important for two reasons:

- non terminal breaks are less relevant in terms of perception; therefore, the fact that almost all problems in segmentation, after only one revision, were related to them means that the original segmentation and the first revision had been accurate;
- non-terminal breaks are precisely the prosodic breaks that relate to the realization of complex informational patterns in utterances, as well as the formation of stanzas and illocutionary patterns.

The proportion of changes according to each type can be better observed in Figure 8. Black slices indicate changes that originate terminal prosodic breaks, gray slices indicate the proportion of changes that create non-terminal breaks, and the hatched portions indicate changes that originate prosodic breaks with no informational value, i.e., retractings and interruptions.

It is clear that the insertion of non-terminal breaks and the switching of terminal breaks to non-terminal breaks are the major changes that must be understood. That is possible if we cross-tabulate the data of these two variables with the information tag that was assigned to them.

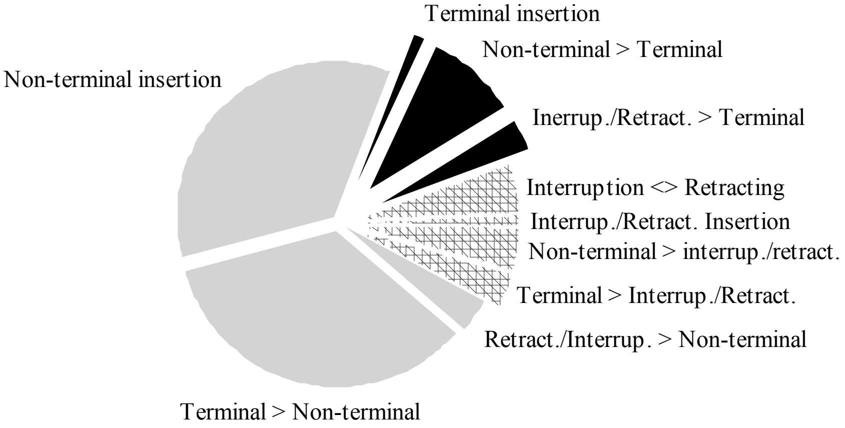

Figure 8. Proportion of different types of changes in prosodic annotation during informational tagging

Table 9 shows the total number of occurrences for each informational tag used in the informationally tagged Brazilian mini-corpus, the total number of changes in prosodic annotation associated with each tag and, also for each tag, the more detailed number of non-terminal breaks insertions and terminal to non-terminal breaks switchings.

These data allow us to see that most switches from terminal to non-terminal break concern the identification of Multiple Comments (CMM) forming illocutionary patterns and Bound Comments (COB) that form stanzas. It is, in fact, difficult sometimes to interpret the value of the prosodic break in cases like these, particularly during the transcription phase, but also during the revision of transcripts that are not aligned with the corresponding audio.

The terminal to non-terminal switching related to COB units reveals that the text-to-speech alignment improves the ability to make refined distinctions about prosodic break values. The annotator can more easily distinguish sequences of units with weak illocutionary value (stanzas) from those that really bear a conclusive prosodic value.

Also the recognition of many illocutionary patterns are facilitated by text-to-speech alignment. In many cases, each root unit (CMM) that composes the illocutionary pattern seems to function in isolation. During the informational tagging, text-to-speech alignment allows the annotator to have the perception of the rhetorical effect created by the units when considered together as part of a unique compound illocutionary pattern. Probably, most of the cases of recognition of illocutionary patterns need the cognitive perspective provided by informational tagging.

Table 9. Cross-tabulation between information tag and change in prosodic breaks annotation

Information tag	**Total tokens**	**Tokens with prosodic annotation changes**		**Non-terminal break insertion**		**Terminal to non-terminal switching**	
COM	4514	166	3.68%	24	14.46%	28	16.87%
CMM	1095	161	14.70%	55	34.16%	**91**	**56.52%**
COB	836	204	24.40%	57	27.94%	**136**	**66.67%**
TOP	581	132	22.72%	**106**	**80.30%**	9	6.82%
EMP	877	86	9.81%	0	0.00%	0	0.00%
SCA	914	79	8.64%	44	55.70%	0	0.00%
PHA	461	47	10.20%	9	19.15%	32	68.09%
PAR	152	30	19.74%	9	30.00%	16	53.33%
INT	236	24	10.17%	7	29.17%	12	50.00%
DCT	177	21	11.86%	17	80.95%	0	0.00%
INP	103	15	14.56%	9	60.00%	5	33.33%
EXP	141	9	6.38%	5	55.56%	4	44.44%
CNT	71	9	12.68%	1	11.11%	8	88.89%
TMT	139	6	4.32%	1	16.67%	1	16.67%
ALL	140	5	3.57%	0	0.00%	4	80.00%
APC	117	5	4.27%	0	0.00%	3	60.00%
APT	23	4	17.39%	4	100.00%	0	0.00%
UNC	53	2	3.77%	0	0.00%	0	0.00%
TPL	22	2	9.09%	2	100.00%	0	0.00%
i-COB	13	2	15.38%	2	100.00%	0	0.00%
i-COM	20	1	5.00%	1	100.00%	0	0.00%
PRL	6	1	16.67%	0	0.00%	1	100.00%
i-CMM	2	1	50.00%	1	100.00%	0	0.00%
i-TPL	1	1	100.00%	1	100.00%	0	0.00%
i-TOP	2	0	0.00%	0	0.00%	0	0.00%
i-PAR	1	0	0.00%	0	0.00%	0	0.00%
Total	10697	1013	9.47%	355	35.04%	350	34.55%

On the other hand, most of non-terminal insertions are linked to the identification of Topic units (TOP). Although this cases are more unexpected and difficult to explain, since Topics are signaled, in principle, with prosodic boundaries of high perceptual salience, two hypotheses can be raised to try to understand why transcribers did not perceive so many prosodic boundaries.

The first one has to do with the fact that a new prosodic profile of Topic was identified during the informational tagging. It is possible that the transcriber's perception was, to some extent, biased by the types of prosodic movements they expected to find. Thus, an unforeseen prosodic movement may have caused the transcribers to disregard it as a prosodic boundary signal. The second hypothesis is that transcribers may have missed non-terminal breaks associated with the border of Topic units when Topics coincide with the subject of the sentence. It is usual that the subject is produced with some prosodic prominence that signals its semantic prominence. Topics, differently, have a prosodic focus that signals its pragmatic prominence, that is to instantiate a cognitive reference for the interpretation of the speech act. It is possible that less experienced transcribers may interpret a Topic as a subject and then miss to annotate the prosodic boundary. Anyway, this is a case that needs further research.

7. Final remarks

This paper presented for the first time two comparable mini-corpora for cross-linguistic analysis of information structure. The two compared languages are Brazilian Portuguese and Italian.

Giving only an overall look to the informational characteristics, it was possible to note some aspects that seem to be language independent, like the basic structure of the three different textual typologies, and some characteristics vary according to the language. A very important difference seems to be the tendency of Brazilian Portuguese to use much less textual units and to be more actional and less textual than Italian. At the same time, we observed that one textual unit, the locutive introducer, is much more used in Brazilian; we proposed an hypothis that could account for this particular feature and that would confirm the general characteristics observed for the different language strategies.

Another important aspect that the two comparable mini-corpora allows us to observe is the completely different behavior of the two languages with respect to dialogic units. These units are a very important feature to study sociolinguistic differences in cross-linguistic verbal behavior.

The last part of the paper aims to show how a different perspective (cognitive versus perceptual) can change the segmentation of the speech flow. The finding of this part of the study can have methodological consequences in speech segmentation, and can help to understand what is more or less salient for perception.

References

Corsi, G. 2009. L'introduttore locutivo: una ricerca corpus-based di italiano parlato informale. BA thesis, Università degli Studi di Firenze.

Cresti, E. & Moneglia, M. (eds) 2005. *C-ORAL-ROM: integrated reference corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: 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.

Cresti, E. 1987. L'articolazione dell´informazione nel parlato. In *Gli italiani parlati: sondaggi sopra la lingua di oggi*. Firenze: Accademia della Crusca, 27-90.

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

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

Cresti, E. 2006. LABLITA Corpus of Spontaneous Spoken Italian. http://lablita.dit.unifi.it/corpora/descriptions/lablita/

Cresti, E. 2009. Unità di analisi testuale e caratteri costruttivi nell'italiano parlato (spontaneo) e scritto (letterario). Ricerche corpus-based. In A. Ferrari (ed.) *Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione, giustapposizione*. Firenze, Cesati, vol. 2, 713-732.

Cresti, E., Panunzi, A. & Scarano, A. 2005. The Italian corpus. In E. Cresti & M. Moneglia (eds), *C-ORAL-ROM: integrated reference corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins, 71-110.

Firenzuoli, V. 2003. Le Forme Intonative di Valore Illocutivo dell'Italiano Parlato: Analisi Sperimentale di un Corpus di Parlato Spontaneo (LABLITA). PhD diss., Università di Firenze.

Firenzuoli, V. & Signorini, S. 2003. L'unità informativa di topic: correlati intonativi. In G. Marotta (ed.), *La coarticolazione: Atti delle XIII giornate di studio del Gruppo di Fonetica Sperimentale*. Pisa: ETS, 177-184.

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

Frosali, F. 2008. Le unità di informazione di ausilio dialogico: valori percentuali, caratteri intonativi, lessicali e morfo-sintattici in un corpus di italiano parlato (C-ORAL-ROM). In E. Cresti (ed.), *Prospettive nello studio del lessico italiano*. Firenze University Press, 417-424.

't Hart, J., Collier R. & Cohen, A. 1990. *A Perceptual Study on Intonation*: *An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press.

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

Maia Rocha, B. 2010. A Unidade Informacional de Introdutor Locutivo no Português Brasileiro: uma análise baseada em corpus. BA thesis, Faculdade de Letras, Universidade Federal de Minas Gerais.
http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/DAJR-8ELJXZ

Maia Rocha, B. & Raso, T. 2011. A unidade informacional de Introdutor Locutivo no português do Brasil: uma primeira descrição baseada em corpus. In *Domínios de Linguagem* 5(1): 1-16.
http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/12479

Martin, P. 2005. WinPitch Corpus: a text-to-speech analysis ans alignment tool. In Cresti, E. & Moneglia, M. (eds), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins, 40-51.

Mello, H., Raso, T., Mittmann, M., Vale, H. & Côrtes, P. In Press. Transcrição e segmentação prosódica do *corpus* C-ORAL-BRASIL: critérios de implementação e validação. In T. Raso & H. Mello (eds) In Press.

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

Moneglia, M. 2011. Spoken Corpora and Pragmatics. *Revista Brasileira de Linguística Aplicada* 11(2): 479-519. http://www.periodicos.letras.ufmg.br/rbla/arquivos/335.pdf

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. & Cresti, E. 2006. C-ORAL-ROM. Prosodic boundaries for spontaneous speech analysis. In Y. Kawaguchi, S. Zaima & T. Takagaki (eds), *Spoken Language Corpus and Linguistics Informatics*. Amsterdam: John Benjamins, 89-112.

Panunzi, A. & Gregori, L. 2011. IPIC: an XML data base for the study of Informational Patterning. Presented at 7th LABLITA International Workshop in Corpus Linguistics, Firenze, http://lablita.dit.unifi.it/lablita_workshop/ws7

Raso, T. In Press. O C-ORAL-BRASIL e a Teoria da Língua em Ato. In T. Raso & H. Mello (eds). In Press.

Raso, T. & Leite, F. 2010. Estudo contrastivo do uso de Alocutivos em italiano, português e espanhol europeus e português brasileiro. *Domínios de Lingu@gem* 4(1): 151-174. http://www.dominiosdelinguagem.org.br/pdf/dl7/DL7-10.pdf

Raso, T. & Ulisses, A. 2008. Tópico e Apêndice no português do Brasil: algumas considerações. *Revista de Estudos da Linguagem* 16(1): 247-262. http://relin.letras.ufmg.br/revista/upload/11-Tommaso_Raso.pdf

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

Raso, T. & Mello, H. 2010. The C-ORAL-BRASIL corpus. In M. Moneglia & A. Panunzi (eds), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Università degli studi di Firenze, 193-213.

Raso, T. & Mello, H. in press. *C-ORAL-BRASIL I. Corpus de referência para a fala espontânea informal do português do Brasil*. Belo Horizonte: UFMG.

Raso, T. & Mittmann, M. 2009. Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem* 17(2): 73-91. http://relin.letras.ufmg.br/revista/upload/17-2_04.pdf

Raso, T., Moraes, J. & Mittmann, M. (forthcoming). *A Topic Prosodic Form in Brazilian Portuguese*.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org

Scarano, A. 2009. A The prosodic annotation of C-ORAL-ROM and the structure of information in spoken language. In L. Mereu (ed.), *Information structures and its interfaces*. Berlin and New York: Mouton de Gruyter, 51-74.

Signorini, S. 2003. Il Topic: criteri di identificazione e correlati morfosintattici in un corpus di italiano parlato. In F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy (eds), *Il parlato* *italiano. Atti del Convegno Nazionale*. Firenze: Franco Cesati, 227-238. http://lablita.dit.unifi.it/preprint/preprint-03coll05.pdf

Tucci, E. 2006. L'unità di appendice in un corpus di italiano parlato (C-ORAL-ROM): caracteristiche intonative, semantiche e morfo-sintattiche. BA thesis, Facoltà di lettere e filosofia, Universitá degli studi di Firenze.

Tucci, I. 2004. L'inciso: caratteristiche morfosintattiche e intonative in un corpus di riferimento. In F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy (eds), *Il parlato italiano. Atti del Convegno Nazionale*. Napoli: D'Auria Editore, 1-14.

Tucci, I. 2009. "Obiter dictum": la funzione informativa delle unità parentetiche. In *Atti del Convegno Internazionale G.S.C.P. "La comunicazione parlata"*. Napoli.

Ulisses, A. 2008. A unidade de Apêndice no português do Brasil. BA thesis, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.