Dialogic units in spoken Brazilian and Italian: A corpus based approach

Maryualê Malvessi Mittmann, Tommaso Raso, Adriellen Arruda

Faculdade de Letras – Programa de Pós-graduação em Estudos Linguísticos

Universidade Federal de Minas Gerais – Belo Horizonte – MG – Brazil

1 INTRODUCTION

In this paper we present a cross-linguistic study on the usage of dialogic units in spoken Brazilian Portuguese and Italian. The data come from two comparable spontaneous speech corpora and the analysis is based on the Language Into Act Theory (Cresti 2000; Cresti and Moneglia 2010), that defines dialogic units as information units (IU) dedicated to regulate the communication. Such units mostly correlate to what other theoretical approaches commonly call discourse markers.

Our main goals are to discuss some interesting aspects regarding the usage of dialogic units in Brazilian Portuguese and Italian. We describe the functions performed by these units in spoken language and as well as present the most frequent lexical items associated to different dialogic functions in Brazilian Portuguese and Italian. We also investigate the distribution of dialogic units across the utterance in order to detect eventual language specific usages.

First, we discuss the notion of discourse markers according to different linguists that have been developing usage based research. Then, we introduce the main concepts of Language Into Act Theory and detail the types and functions of information units in this approach. Research methods are described in the following section. The results' section contains quantitative data and a few qualitative analyses, showing examples extracted from the corpora. All examples presented in the text can be obtained through IPIC database (IPIC 2012), where the audio signal is available together with text transcription and recording session metadata.

2 CONCEPTUAL ISSUES AND THEORETICAL FRAMEWORK

2.1 Conceptual issues on discourse markers

In linguistics literature, discourse markers are often defined as linguistic expressions that lose their semantic meaning and its original morphosyntactic value that do not belong to the

semantic and syntactic structure of the utterance. Such expressions do not affect the truth value of the utterance (Scneider 1999), and are not part of the propositional content of the message conveyed, therefore not contributing to the meaning of the proposition itself (Fraser 2006).

According to different traditions, discourse markers acquire different pragmatic functions, which can be either textual or meta-textual. Some textual features usually attributed to discourse markers are turn-taking, silence filling, phatic function, request for attention, agreement and confirmation. Meta-textual functions can be focus, demarcation, indication of paraphrase or reformulation, modality, among others (Fisher 2006). However, one can argue that several of the "textual" functions, such as turn taking or request for attention, are actually pragmatic functions, since they do not contribute with the propositional content of the utterance.

Additionally, there is little agreement in the literature regarding the number of discourse markers, as to their functions and the criteria to define them. Discourse markers are often related to concepts such as form, attitude and emotion (Traugott 2007), but there are also no agreement regarding these concepts. Some authors note a strong correlation between discourse markers and some prosodic properties, such as the fact that they tend to be uttered in a dedicated tone unit that can be eliminated without any effect on the utterance (Bazzanella et al. 2008).

2.2 Language Into Act Theory

The theoretical framework adopted here, Language Into Act Theory (Cresti 2000; Cresti and Moneglia 2010; Cresti 2011), developed through empirical corpus research that identifies such expressions as dialogic information units. Dialogic units are prosodically delimited linguistic expressions that function regulating the communicative interaction (Cresti 2000; Frosali 2008).

In Language into Act Theory, the referring unit for the analysis of the spoken language is the utterance, defined as the linguistic counterpart of a speech act. The utterance is the shortest linguistic unit that can be pragmatically interpreted and is delimited in the speech flow by prosodic boundaries that bear a conclusive value. In spontaneous speech, prosody plays a fundamental role of parsing the speech flow into discrete tone units. Tone units can be prosodically autonomous or prosodically non-autonomous. Autonomous tone units are those delimited by prosodic boundaries perceived by the hearer as having a conclusive value. Prosodically delimitated linguistic sequences - tone units - convey information. The information is pragmatically autonomous if it bears an illocutive value. Tone units that convey illocution are both prosodically and pragmatically autonomous and are associated with the Comment function. Tone units that are delimited by prosodic boundaries with a non-conclusive value convey other types of information and are associated with different functions.

The utterance may be organized in a single tone unit (simple utterance) or it can be prosodically parsed into two or more tone units (compound utterance), creating a prosodic pattern (Hart, Collier, and Cohen 1990). The units of a prosodic pattern are associated with information functions, through which information is patterned in the utterance.

Informational Patterning Hypothesis (Cresti and Moneglia 2010; Scarano 2009) proposes that there is a systematic correspondence between the prosodic pattern and the information pattern of an utterance. Information Units (IU) are classified into textual and dialogic. Textual units participate to the construction of the semantic content of the utterance. Dialogic units are devoted to the successful pragmatic performance of the utterance (e.g. to regulate the relationship between speakers). Every utterance has at least one Comment unit, since it is the Comment that bears the utterance's illocutionary force. The Comment is the only necessary and sufficient unit to form an utterance.

Textual functions are:

(a) Topic: Identifies the domain of application for the illocution;

(b) Appendix of comment: Integrates the text of the comment;

(c) Appendix of topic: Integration of the information given in the topic;

(d) Parenthesis: Adds information with metalinguistic value;

(e) Locutive introducer: Signals a change of point of view on the subsequent locution.

The dialogic functions are:

(a) Incipit: Opens the communicative channel while signals a contrastive value with the previous utterance;

(b) Conative: Pushes the listener to take part in an adequate way in the dialogue;

(c) Phatic: Ensures the maintenance of the communicative channel;

(d) Allocutive: Specifies to whom the message is directed;

(e) Expressive: Emotional support of the utterance;

(f) Discourse Connector: Signals the continuity of the discourse while establishes a relation between the previous and following units.

3 METHODS

We present two samples of spoken corpora that received tagging at the information structure level according to the Language into Act Theory. The Italian sample comes from the C-ORAL-ROM (Cresti and Moneglia 2005) (Italian section) and the Brazilian sample comes from C-ORAL-BRASIL (Raso and Mello 2012).

The samples come from informal sections of oral corpora containing a broad variety of communicative situations and were selected for a strict comparison with each other. The corpora are representative of spontaneous speech, recorded in natural, not controlled, communicative situations. Recordings are transcribed in CHAT format (MacWhinney 2000) with annotation of prosodic boundaries (Moneglia and Cresti 1997). The prosodic boundary annotation was validated in both corpora (Moneglia, Scarano, and Spinu 2005; Raso and Mittmann 2009). Box 1 shows the symbols used in the annotation of prosodic phenomena.

Box 1: Prosodic breaks annotation scheeme in C-ORAL corpora

Signal	**Meaning**
?	It delimits a prosodically autonomous sequence with a clear interrogative prosodic profile.*
…	It delimits a prosodically autonomous sequence voluntarily interrupted by the speaker with a suspensive prosodic profile.*
+	It signals unintentionally interrupted sequences. In this case, the speaker's program is broken and the interpretability of the sequence can be compromised
//	It indicates a terminal break, marks all prosodically autonomous sequences that do not belong to the previous classes.
/	It signals non terminal prosodic breaks, it delimits TU.
[/n]	It 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 involved in false starts do not contribute to the informational patterning and to the semantic content of the Utterance

* Used only in C-ORAL-ROM. For these cases, C-ORAL-BRASIL uses the // sign.

The Italian sample[1] contains 29414 words, 5286 utterances and 11517 prosodic/information units. The Brazilian Portuguese sample[2] has 31318 words, 5483 utterances and 9825 prosodic/information units. Samples are balanced for type of communicative interaction (see Table 1).

Table 1: Number of recorded sessions according to communicative context and type of interaction

Corpus section		Number of sessions	
Communicative context	Type of interaction	Brazilian	Italian
Family/ Private	Monologue	6	5
	Dialogue	5	6
	Conversation	4	3
Public	Monologue	1	2
	Dialogue	2	2
	conversation	2	2
Total		20	20

[1] The Italian sample is identified in the online database (IPIC) as *MiniCorpus_Ita.*
[2] The Brazilian Portuguese sample is identified in the online database as *Brasiliano.*

Samples received manual annotation of information functions, according to the information units proposed by Language into Act Theory. We extracted the data through IPIC (first release), a theoretically-bound XML Database designed for the study of linear relation among Informative Units in spoken language corpora (Panunzi and Gregori 2011). The tagset is available at the database website (Anon. 2012). The data were tabulated and we analyzed the frequencies and distribution of all information units in both samples.

4 RESULTS

According to Language into Act framework, pragmatics operates at two levels:

(1) The macro-pragmatic level is related to the production of Speech Acts (Austin 1962). It organizes the speech flow into pragmatically autonomous linguistic sequences (utterances).

(2) The micro-pragmatic level is related to the patterning of information within the utterance. It organizes utterances into patterns of information units (IU).

Results show a prevalence of compound utterances in Italian (36%) in comparison with Brazilian (29%) that is statistically significant (chi-square=52,848 – $p<0.0001$). That means that Italian presents a higher frequency of utterances formed by the Comment unit plus one or more information units with different functions. Italian is then much more structured at the micro-pragmatic level than Brazilian Portuguese.

In principle, that increases the probability of Italian having a higher frequency of Dialogic Units. Interestingly, in Italian, information is more likely patterned at the textual level, with high occurrence of compound Utterances with only textual IU (43% of all compound Utterances). On the contrary, Brazilian presents a more frequent use of dialogic IU (50% of all compound Utterances). Table 2 shows the frequency of different types of compound information structures in Brazilian Portuguese and Italian.

Table 2: Distribution of different types of compound utterances

Information Structure	BP				Italian			
	cv.	dl.	mn.	tot	cv.	dl.	mn.	tot
Compound utterances with only dialogic units	20%	24%	6%	50%	11%	14%	9%	35%
Compound utterances with only textual units	11%	13%	11%	35%	14%	14%	15%	43%
Mixed compound utterances	4%	5%	5%	14%	6%	6%	9%	22%
Total	35%	42%	23%	100%	32%	34%	33%	100%

The use of dialogic units also differs among Brazilian and Italian, as showed in Figures 1 and 2.

Frequency of dialogic units in Brazilian Portuguese and Italian

Dialogic unit	Brazilian Portuguese	Italian
Phatic	42%	46%
Discourse conector	16%	9%
Expressive	13%	3%
Allocutive	13%	5%
Incipit	9%	29%
Conative	7%	8%

Comparing Brazilian and Italian with respect to all the dialogic units, we note that Brazilian uses much more Expressives and Allocutives, while Italian uses much more Incipits.

The Incipit primary pragmatic function is to take the turn, expressing a strong opposition regarding the previous utterance. In Brazilian Portuguese, taking the turn with an Incipit can sound rude. Incipits are not only much more frequent in Italian, but they also present a greater lexical variety, with a type/token ratio of 0,11 (46 types and 411 tokens). Typical Incipits in Italian are "allora" (so), "però" (but) and "no" (no). Brazilian, on the other hand, has a type/token ration of Incipits of 0,13 (14 types and 104 tokens). The most frequent lexical items with Incipit function are "não" (no), "é" (yes) and "ah" (interjection). See examples (1) and (2).

(1) *BAL: não /=INP= mas é porque eu tô pensando assim //=COM=
no / but I'm thinking like this //

(2) *MAX: allora /=INP= entriamo /=CMM= e facciamo la benzina /=CMM= vai //=CNT=
so / lets go in [the gas station] /and put the gas / c'mon //

Expressives are not very frequent in Italian, which has a type/token ratio of 0,41 (20 types and 48 tokens. In Brazilian, Expressives show a token/type ratio of 0,18 (26 types and 141 tokens). Typical Expressives are interjections (ah, eh) and also expressions of religious character, like the Italian Madonna (Mother of God) "Nossa" (a short for "Our Lady"). Expressives convey an emotional value associated with the speech act. In Italian, the primary function is search of social cohesion through speech. In Brazilian, this function conflates with the turn taking function. Expressives provide a softer, more polite way to Brazilians open the utterance and/or to take the turn. See examples (3) and (4).

(3) *PAU: ah /=EXP= eu tenho uma aqui //=COM=

ah / I have one here //

(4) *ELA: eh /=EXP= birbone hhh //=COM=

eh / naughty boy [hhh=laughter] //

Allocutives are also not very frequent in Italian and have a type/token ratio of 0,18 (12 types for 67 tokens). Italians use Allocutives mostly to identify the messages addressee, and therefore they are very frequent in conversations among three or more people and very rare in dialogues and monologues. Brazilian Portuguese has a really high frequency of Allocutives, with a type/token ratio of 0,13 (18 types and 140 tokens). Lexical items that are commonly used with this function are proper nouns, nicknames and expressions such as “ragazzi” (guys) in Italian or “filho/filha” (son/daughter) in Brazilian Portuguese. See examples (5) and (6).

(5) *CAR: é o quatro mesmo / Jacaré //=ALL=

it's the four indeed / Jacaré *[=Alligator] //*

(6) *ELA: e te / Massimo /=ALL= quanto tu <c' avevi> ?

and you / Massimo / how much did you have ?

Brazilians do a more frequent use of Allocutives to demonstrate social cohesion towards the interlocutor. In Brazilian, they are frequent not only in conversations (as in Italian), but also in dialogues and monologues.

When we look at the distribution of dialogic units regarding its position inside the utterance, we notice that the Expressives are very often employed to open the utterance and/or to take the turn. In Italian, those functions are mostly performed by Incipits. Allocutives and Expressives are signs of social cohesion in discourse, while Incipits signal the speaker's opposition with respect to the previous utterance. It is likely that in Brazilian culture the Incipit is perceived as an aggressive way to take the turn or begin the utterance. For this reason, Brazilian tends to prefer Expressives to play this role.

5 FINAL REMARKS

Taking the acoustic signal and the natural prosodic parsing of speech into account provides

valuable tools for the linguistic analysis, since allows to properly interpreting the pragmatic value of information units that compose the utterance. Distribution alone is not sufficient to the analysis of spoken data. Without the acoustic information, it is impossible to disambiguate the function of one same lexical item in a given position in different utterances.

The differences observed in Italian and Brazilian Portuguese regarding the use and distribution of Dialogic Units suggest cultural influences in language use. Dialogic units are strongly linked to the interaction (and not the semantic content of the utterance) and therefore, sensitive to cultural nuances.

Cross-linguistic studies are very valuable, because through the analysis of different languages we can observe which features are intrinsic to speech as a universal communicative medium and which ones are specific of each language. Individualizing what is specific to each language is necessary to develop and implement appropriate teaching strategies, translation tools and better NLP systems. The presence of comparable corpora and the study of the information structure in a contrastive perspective provide many useful elements for L2 teaching. It is clear that the pragmatic perspective, often invoked in education, still lacks appropriate tools of research. Corpora such as C-ORAL-ROM and the C-ORAL-BRASIL and a theoretical perspective as Language into Act Theory can provide tools to repair this deficiency.

REFERENCES

IPIC (2012). "IPIC: Information Structure Database." *DB IPIC First Release*. http://lablita.dit.unifi.it/app/dbipic/.

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

BAZZANELLA, C, et al. (2008). "Polifunzionalità Dei Segnali Discorsivi, Sviluppo Conversazionale e Ruolo Dei Tratti Fonetici e Fonologici", *in* PETTORINO, M. et al. (ed.), *La Comunicazione Parlata*. Napoli: Liguori, pp. 934–963.

CRESTI, E. (2000). *Corpus Di Italiano Parlato*, vol. 1. Firenze: Accademia della Crusca.

——— (2011). "The Definition of Focus in Language into Act Theory.", *in* MELLO, H., PANUNZI, A. and RASO, T. (ed.), *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*. Firenze: FUP, pp. 39–82.

CRESTI, E. and MONEGLIA, M. (eds.) (2005). *C-Oral-Rom: Integrated Reference Corpora For Spoken Romance Languages.* Amsterdam: John Benjamins.

CRESTI, E. and MONEGLIA, M. (2010). "Informational Patterning Theory and the Corpus-based Description of Spoken Language: The Compositionality Issue in the Topic-comment Pattern", *in* MONEGLIA, M. and PANUNZI, A. (eds.), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze: FUP, pp. 13–45.

FISHER, K. (2006). “Towards an Understanding of the Spectrum of Approaches to Discurse Particles”, *in* FISHER, K. (ed.), *Approaches to Discourse Particles*. Amsterdam: Elsevier, pp. 1–20.

FRASER, B. (2006). “Towards a Theory of Discourse Markers”, *in* FISHER, K. (ed.), Approaches *to Discourse Particles*. Amsterdam: Elsevier, pp. 189–204.

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* CRESTI, E. (ed.), *Prospettive Nello Studio Del Lessico Italiano.* Firenze: FUP, pp. 417–424.

HART, J.’t, COLLIER, R and COHEN, A. (1990). *A Perceptual Study on Intonation: An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press.

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

MONEGLIA, M. and CRESTI, E. (1997). “L´intonazione e i Criteri Di Trascrizione Del Parlato Adulto e Infantile”, *in* BORTOLINI, U. and PIZZUTO, E. (ed.), *Il Progetto CHILDES Italia.* Pisa: Del Cerro, pp. 57–90.

MONEGLIA, M., SCARANO, A. and SPINU, M. (2005). “The C-ORAL-ROM Multilingual Corpus of Spontaneous Speech: Validation of the Prosodic Annotation by Expert Transcribers”, *in* MARTÍNEZ, C. N. and MONELGIA, M. (eds.). *Computers, Literature and Philosophy CLIP 2003*.Firenze: FUP, pp. 107–120.

PANUNZI, A. and GREGORI, L. (2011). “DB-IPIC: an xml database for the representation of information structure in spoken language”, *in* MELLO, H., PANUNZI, A. and RASO, T. (ed.), *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*. Firenze: FUP, pp. 133–150

RASO, T. and MELLO, H. (eds.) (2012). *C-ORAL-BRASIL I: Corpus De Referência Do Português Brasileiro Falado Informal*. Belo Horizonte: UFMG.

RASO, T. and MITTMANN, M. M. (2009). “Validação Estatística Dos Critérios De Segmentação Da Fala Espontânea No Corpus C-ORAL-BRASIL.” *Revista De Estudos Da Linguagem, vol 2, nº* 17, pp. 73–91.

SCARANO, A. (2009). “A The Prosodic Annotation of C-ORAL-ROM and the Structure of Information in Spoken Language”, *in* MEREU, L. (ed.), *Information Structures and Its Interfaces*. Berlin: Mouton de Gruyter, pp. 51–74.

SCHNEIDER, S. (1999). *Il Congiuntivo Tra Modalita e Subordinazione: Uno Studio Sull’italiano Parlato*. Roma: Carocci.

TRAUGOTT, E. (2007). “Discourse Markers, Modal Particles, and Contrastive Analysis, Synchronic and Diachronic.” *Catalan Journal of Linguistics, nº* 6, pp. 139–157.