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

Maryualê M. MITTMANN; Tommaso RASO; Adriellen ARRUDA

Universidade Federal de Minas Gerais

Summary

1. Dialogic units or discourse markers?
   Some theoretical discussion

2. Does prosody matter?
   The LAct approach

3. Spontaneous speech:
   Tagging and extracting data

4. Information structure or Lexicon?
   Italian vs Brazilian information strategies

5. And know what?
   Possible applications

Dialogic units or discourse markers?

- **Discourse markers:**

  - Lost of original semantic meaning and morphosyntactic value;

  - Do not partake of the semantics and syntax of the utterance;

    - ⚝ Free distribution

    - ⚝ Pragmatic functions

Dialogic units or discourse markers?

- **Discourse markers:**

  - Shoroup (1999): optional connecting expressions, that do not affect the truth value of the utterance.

  - Fischer (2006): DM functions can be...

    - **Textual:** turn-taking, silence filling, phatic, attention request, agreement, confirmation.

    - **Meta-textual**: focus, demarcation, indication of reformulation, modality.

Dialogic units or discourse markers?

- **Discourse Markers - DM:**

  - Traugott (2007): expression of modality, attitude and emotion.

    - No agreement regarding those concepts! (See Mello & Raso 2012).

  - Bazzanella et al. (2008): correlation between discourse markers and prosody.

    - DM tend to be uttered in a dedicated prosodic unit.

Does prosody matter?

LAct – Language into Act Theory

▶ Prosodic boundaries delimit linguistic sequences:

- Prosodically autonomous (concluded)

- Prosodically non-autonomous (not concluded)
  → Prosodic Pattern (Hart, Cohen & Collier, 1990)

▶ Prosodically delimitated linguistic sequences – prosodic or tone units - convey information

Does prosody matter?

LAct – Language into Act Theory

▶ Prosodic units convey information:

- Pragmatically autonomous:
Ex. (1): REN: uai //

  ✴ Illocutionary force = Comment IU.

- Pragmatically non-autonomous:
Ex. (2): HEL: uai / cê pode fazer assim / mas cê nũ pode fazer assim //

  ✴ No illocution = other type of IU.

Does prosody matter?

LAct – Language into Act Theory

▶ Prosodic units convey information:

- **No relation with propositional content!**

Ex. (3): BEL: pois é //

Ex. (4): BAL: porque / <se eu for> empregado / por exemplo / alguém vê que eu sou muito foda / <medo> de perder / <o posto> <deles / es vão [/2] es vão> me dizer / <né> //

Does prosody matter?

LAct – Language into Act Theory

▶ Utterance: shortest linguistic unit that can be pragmatically interpreted → Speech Act.

- Simple: single prosodic/information unit.

- Compound: two or more prosodic units.
  → Information Pattern (Cresti, 2000).

Does prosody matter?

▶ Information Units (IU) can be textual or dialogic.

- **Textual:** construction of the semantic content of the utterance.

  - ✶ Comment: nuclear IU, illocutionary value.

- **Dialogic:** success of pragmatic performance of the utterance. → *Discourse Markers!*

Ex. (5): DUD: pô / Mailton / eu nũ entendo muito de cobra não / mas essa história daí / eu acho que quem matou o cara foi a mulher dele / hein //

Does prosody matter?

▶ Textual IU with no illocutionary value:

- Topic – identifies the domain of application for the illocution;

- Appendix – integrates the text of the Comment or Topic;

- Parenthesis – adds information with metalinguistic value;

- Locutive Introducer – signals a change of point of view on the subsequent locution.

Does prosody matter?

- Dialogic IU:

  - Incipit – opens the communicative channel while signals a contrastive value with the previous utterance;

  - Conative – pushes the listener to take part in an adequate way in the dialogue;

  - Phatic – ensures the maintenance of the communicative channel;

Does prosody matter?

- Dialogic IU:

  - Allocutive – specifies to whom the message is directed; signals social cohesion;

  - Expressive - emotional support of the utterance;

  - Discourse Connector – signals the continuity of the discourse while establishes a relation between the previous and following units.

Spontaneous speech:
Tagging and extracting information

▶ C-ORAL-ROM IT and C-ORAL-BRASIL:

- Transcription and annotation of prosodic boundaries.

Ex. (6): MAI: e &di e existe uma cobra lá naquele interior que é muito muito enorme de grande eu nũ sei o nome dela muito grande

MAI: e &di [/2] e existe uma cobra / lá naquele interior / que é muito [/1] muito enorme de grande / eu nũ sei o nome dela // muito grande //

Spontaneous speech:
Tagging and extracting information

▶ Mini-corpora IT and BP:

- Tagging: Association of information function to each prosodic unit.

Ex. (6'): MAI: e &di [/2]=EMP= e existe uma
cobra /=i-COB= lá naquele interior /=PAR=
que é muito [/1]=SCA= muito enorme de
grande /=COB= eu nũ sei o nome
dela //=COM= muito grande //=COM=

Spontaneous speech:
Tagging and extracting information

▶ The IT sample (*Minicorpus Italiano*):

  - 29414 words

  - 5286 utterances

  - 11517 prosodic/information units.

▶ The BP sample *(Minicorpus Brasileiro)*:

  - 31318 words

  - 5483 utterances

  - 9825 prosodic/information units.

Spontaneous speech:
Tagging and extracting information

▶ Data extraction: IPIC

- Theoretically-bound XML Database.

- Designed for the study of linear relation among Informative Units in spoken language corpora.
  (Panunzi & Gregori, 2012)

http://lablita.dit.unifi.it/ipic/

Information structure or lexicon?

Information structure in IT and BP

▶ Italian: 30% compound utterances

▶ Brazilian: 23% compound utterances
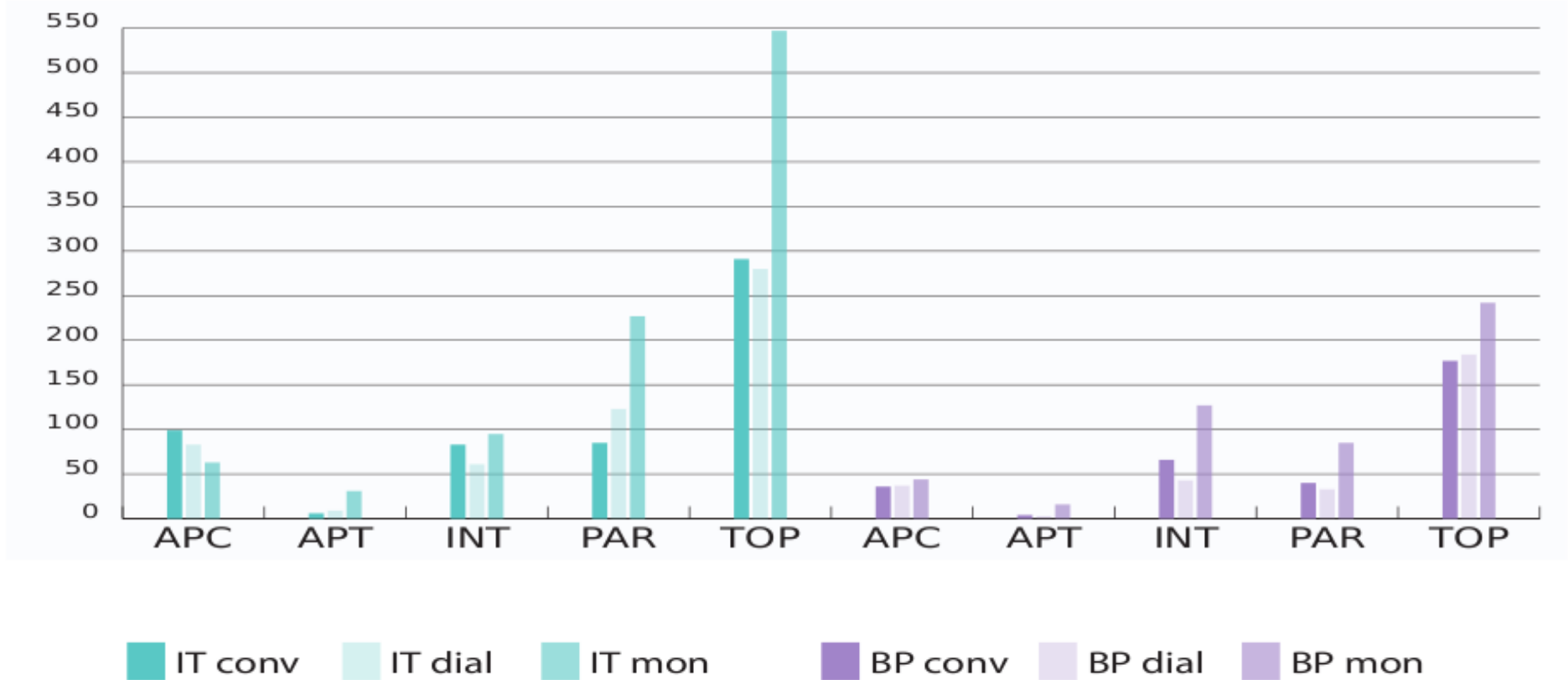
COM + Textual UI ← **IT 44%   51% BP** → **COM + Dialogic UI**

Information structure or lexicon?
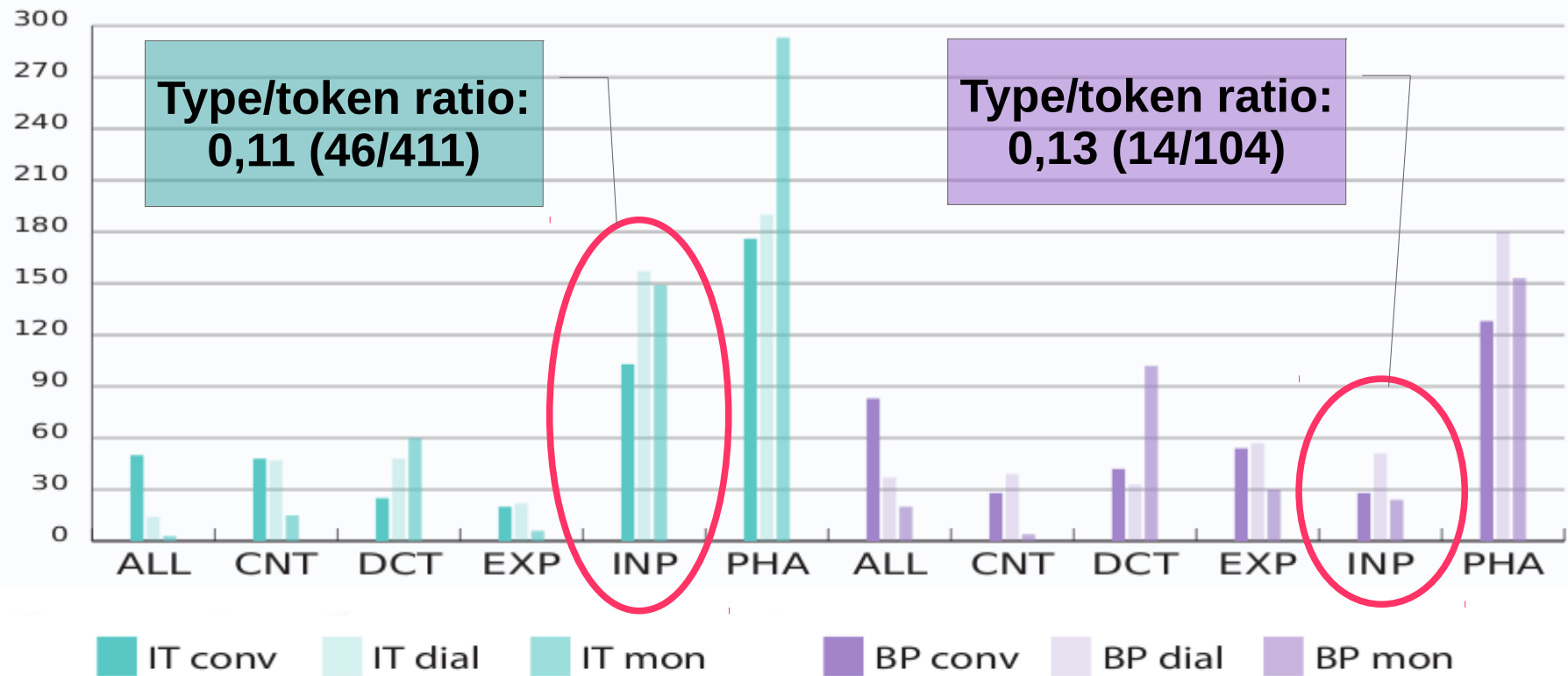
Information structure in IT and BP

Textual units

Legend: IT conv, IT dial, IT mon, BP conv, BP dial, BP mon

Information structure or lexicon?

Information structure in IT and BP: Incipit

Dialogic units

Type/token ratio: 0,11 (46/411)

Type/token ratio: 0,13 (14/104)

Legend: IT conv, IT dial, IT mon, BP conv, BP dial, BP mon

Categories: ALL, CNT, DCT, EXP, INP, PHA

Information *structure* or lexicon?

▶ Use of Incipit: Strong opposition regarding previous utterance.

- BP – Turn taking, but can sound rude.
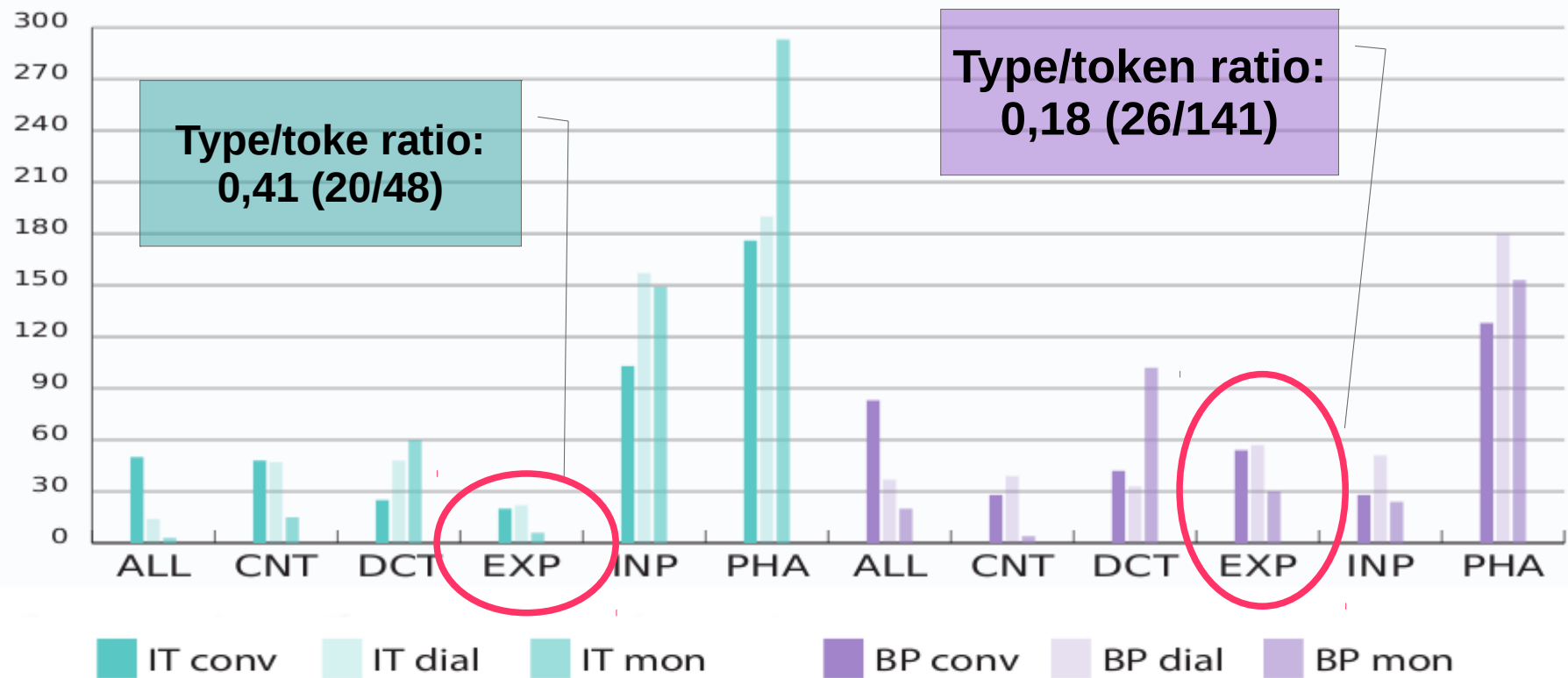
- IT – Turn taking.

Lexical selecion in IT and BP

Ex. (7): BAL: não /=INP= mas é porque eu tô pensando assim //

Ex. (8): MAX: allora /=INP= entriamo / e facciamo la benzina / vai //

Information structure or lexicon?

Information structure in IT and BP: Expressive

Dialogic units

Type/toke ratio:
0,41 (20/48)

Type/token ratio:
0,18 (26/141)

Legend: IT conv, IT dial, IT mon, BP conv, BP dial, BP mon

Information *structure* or lexicon?

▶ Use of Expressive: emotion associated with the speech act.

  - BP - very often employed as a softer way to open the utterance and/or to take the turn.

  - IT - marking social cohesion.

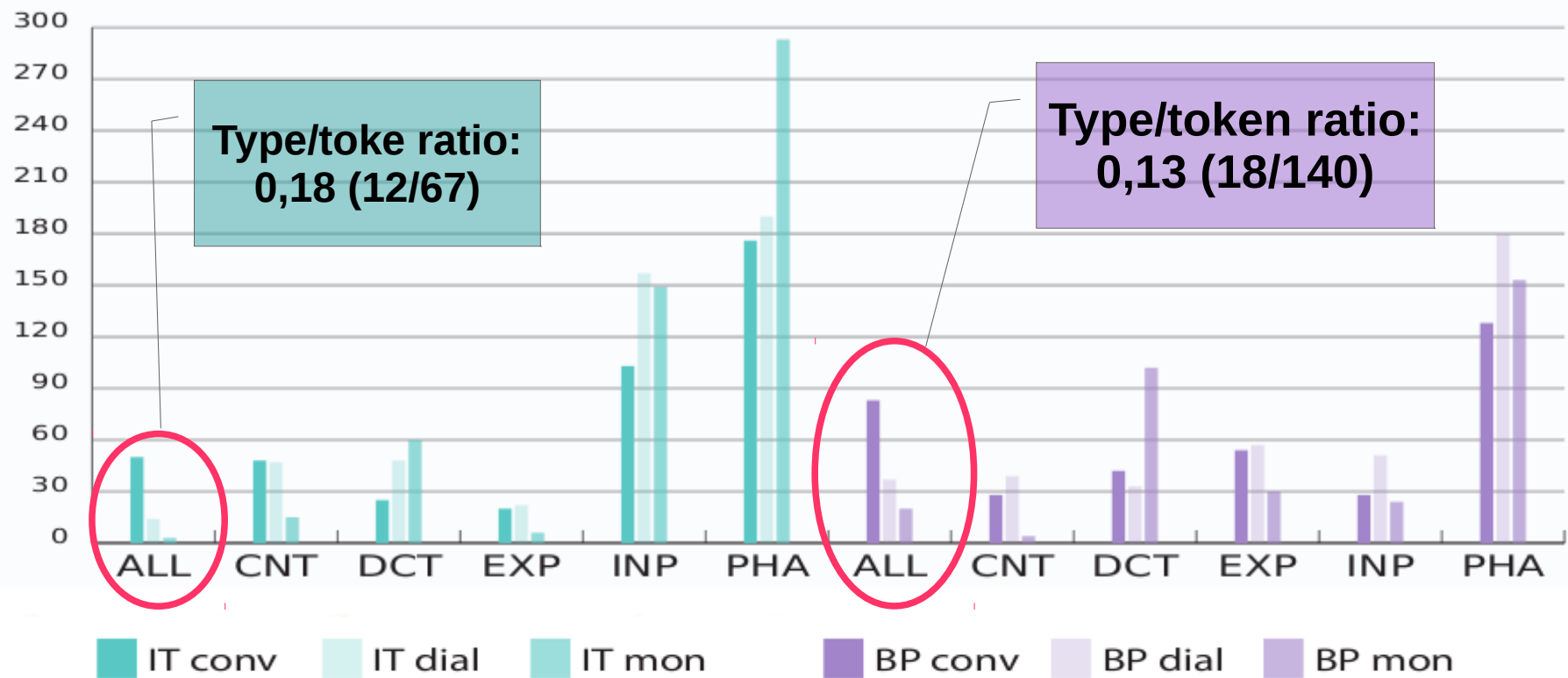Lexical selection in IT and BP: Expressives

Ex. (9): ah /=EXP= eu tenho uma aqui //=COM=

Ex. (10): eh /=EXP= birbone hhh //=COM=

Information structure or lexicon?

Information structure in IT and BP: Allocutive

Dialogic units

Type/toke ratio:
0,18 (12/67)

Type/token ratio:
0,13 (18/140)

IT conv | IT dial | IT mon | BP conv | BP dial | BP mon

Information *structure* or lexicon?

Lexical selecion in IT and BP

▶ Use of Allocutives:

- BP – Social cohesion (high use in dl and mn).

- IT – identify the message's addressee
    (high use in cv).

Ex. (11): CAR: é o quatro mesmo / Jacaré //=ALL=

Ex. (12): ELA: e te / Massimo /=ALL= quanto tu
    <c' avevi> ?

And know what?

- Through the analysis of different languages we can observe:
    - What is intrinsic to speech?
    - What is language specific?

- Same information units may have different distribution, lexical selection, and cultural related communicative nuances depending on the language.
    - Prosody helps us with that!

And know what?

▶ Annotated and aligned spoken corpora make possible to work with "large" amounts of spontaneous speech data.

▶ And then:

- Develop better teaching materials and strategies;

- Help translators and improve their tools;

- Develop more efficient NLP systems.

Acknowledgments

► This work was developed as part of the collaboration agreement between:

- LABLITA - Linguistic Laboratory of the Italianistic Department - University of Florence (UNIFI).

- LEEL – Laboratório de Estudos Empíricos da Linguagem - Federal University of Minas Gerais (UMFG).

► C-ORAL-BRASIL Project is funded by:

U F m G
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

FAPEMIG

References

Bazzanella, C.; Bosco, C.; Gili Fivela, B.; Miecznikowski, J.; Tini Brunozzi, F. (2008) Polifunzionalità dei segnali discorsivi, sviluppo conversazionale e ruolo dei tratti fonetici e fonologici. In: Pettorino, M.; Giannini, A.; Vallone, M.; Savy, R. (Eds.) La comunicazione parlata, vol. II. Napoli: Liguori, p. 934-963.

Cresti, E. (2000) Corpus di italiano parlato. Firenze: Accademia della Crusca.

Cresti, E. (2011) The Definition of Focus in Language into Act Theory (LAcT). In: Mello, H.; Panunzi, A.; Raso, T. Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation. Firenze: FUP. p. 39-82

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: Moneglia, M.; Panunzi, A. (eds). Bootstrapping Information from Corpora in a Cross-Linguistic Perspective. Firenze: FUP.

Cresti, E.; Moneglia, M. (Eds.) (2005) C-ORAL-ROM. Integrated reference corpora for spoken romance languages. Amsterdam: John Benjamins.

References

Fischer, K. (2006) Towards an understanding of the spectrum of approaches to discurse particles. In: Fischer, K. Approaches to discourse particles. Amsterdam: Elsevier, p. 1-20.

Fraser, B. (2006) Towards a Theory of Discourse Markers. In: Fischer, K. (Ed.) Approaches to discourse particles. Amsterdam: Elsevier, p. 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. (Org.) Prospettive nello studio del lessico italiano. Firenze: Firenze University Press, p. 417-424.

Hart, J't.; Collier, R; Cohen, A. (1990) A perceptual study on intonation: An experimental approach to speech melody. Cambridge: Cambridge University Press.

Panunzi, A.; Gregori, L. (2011) DB-IPIC: an XML database for the representation of information structure in spoken language. In: Mello, H.; Panunzi, A.; Raso, T. Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation. Firenze: FUP. p. 133-150

References

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

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.

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 6, p. 139-157.