The LABLITA spoken corpora collection and the Language into Act Theory

Emanuela Cresti, Massimo Moneglia, Alessandro Panunzi
University of Florence, LABLITA

The Language into Act Theory (L-AcT) has been developed in Italy since the eighties and aims at providing a framework suitable for the corpus-based study of spontaneous speech (Cresti 2000, 2018; Cresti & Moneglia 2018; Cresti *et al.* 2018b).

L-AcT was applied to the Italian LABLITA corpus (1.135.000 transcribed words, 107.000 reference units text/sound aligned) (Cresti *et al*. 2018a) and has been heavily tested in the collection and annotation of Romance corpora: C-ORAL-ROM (Cresti & Moneglia 2005), C-ORAL-BRASIL (Raso & Mello 2012), Cor-DiAL (Nicolas Martinez 2013). Transcripts complies with the CHAT - LABLITA Format (Cresti & Moneglia 2005), text / sound alignment and acoustic analysis have been achieved through the software WINPITCH (Martin 2003, 2015). Both the LABLITA corpus and the C-ORAL-ROM Italian collection, for the qualities of their corpus design, constitute reference corpora for Italian.

This framework was also used for grounding the cross-linguistic comparison of Information Structure in spontaneous speech (Moneglia & Raso 2014). To this end the IPIC Data Base was created and applied to Italian and Brazilian Portuguese tagged corpora (Panunzi & Mittmann 2014). The extension of IPIC to Spanish was recently achieved (Nicolas Martinez & Lombán forthcoming). American English has been also tagged according to the same methodology on the basis of a selection of S. Barbara corpus (Du Bois *et al.* 2000; Cavalcante & Ramos 2016).

Within the Austinian tradition, L-AcT assumes that the utterance is the counterpart of a speech act and constitutes the primary reference unit for the analysis of speech. Its main novelty with respect to Austin is to consider that the spoken activity manifests through prosodic devices, specifically for what regards the core levels of Illocutionary force and Information structure (IS). Therefore, the processing of prosody is assumed as a mandatory step for the identification in the flow of speech of both Utterances and their Information Units.

L-AcT foresees the systematic correspondence between stretches of speech ending with a *terminal prosodic break* and the accomplishment of an utterance, and, within the utterance, between chunks segmented by a *non- terminal break* and information functions (Cresti & Moneglia 2005; Izre'el & Mettouchi 2015).

The presentation will sketch the spoken corpora previously mentioned and will focus on the methodology for the detection of prosodic breaks and its validation (Danieli *et al*. 2004; Moneglia *et al*. 2010; Raso & Mittmann 2009).

The relevance of prosodic breaks will be highlighted tracing back to the IPO tradition that assumes that intentionally performed prosodic cues are significant to perception ('t Hart *et al*. 1990; Firenzuoli 2003). Current trends in the L-AcT framework concerning both perceptual and automatic detection of breaks will be also referred (Barbosa & Raso 2018).

Sequences ending with a terminal break strictly correspond to speech Reference Units (Izre'el *et al.* forthcoming) and may correspond to *utterances* matching with one speech act (90% of cases in the above corpora) or to *stanzas*, corresponding to the expression of a flow of thought (Chafe 1970). Utterances and stanzas are the reference entities suitable for the identification of syntactic and semantic relations in speech.

The added value of the annotation of terminal breaks for the use of spontaneous speech data in linguistic research will be considered. Their detection in the acoustic source determines the alignment unit and specifies the higher level of linguistic annotation for parsing the speech flow into information units and syntactic chunks. In other words, the annotation of terminal breaks in the acoustic source determines the reference units and specifies the higher level of linguistic annotation.

REFERENCES

Cavalcante F. & Ramos A. (2016). The American English spontaneous speech minicorpus. Architecture and comparability. *CHIMERA*, 3(2).

Chafe W. (1970). *Meaning and the Structure of Language*. Chicago: CUP.

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

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

Cresti E. & Moneglia M. (2018). The illocutionary basis of Information Structure. Language into Act Theory (L-AcT*)*. In E. Adamou, K. Haude & M. Vanhove (eds), *Information structure in lesser-described languages: Studies in prosody and syntax*. Amsterdam: Benjamins, 359-401.

Cresti E., Moneglia M. & Panunzi A. (2018a). The LABLITA Corpus & the Language into Act Theory: analysis of Viterbo excerpts. In A. De Dominicis (ed.), *Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis*. Contributi del Centro Linceo "Beniamino Segre", vol. 137. Roma: Bardi Edizioni, 47-63.

Cresti E., Gregori L., Moneglia M. & Panunzi A. (2018b). The Language into Act Theory: A Pragmatic Approach to Speech in Real-Life. In H. Koiso & P. Paggio (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), *LB-ILR2018 and MMC2018 Joint Workshop: Language and Body in Real Life* Multimodal Corpora & *Multimodal Data in the Online World*. Paris: ELRA, 20-25.

Danieli et al. (2004). Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech C-ORAL-ROM. In C. Draxler, H. van den Heuvel & F. Schiel (eds), *Proceedings of LREC 2004*. Paris: ELRA, 1513–1516.

Du Bois J., Chafe W., Meyer Ch. & Thompson S. (2000). *Santa Barbara Corpus of Spoken American English Part 1*, LDC2000S85. Philadelphia: Linguistic Data Consortium.

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

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

Izre'el S., Mello H., Panunzi A. & Raso T. (eds) (forthcoming). *In search for a reference unit of spoken language: a corpus driven approach*. Amsterdam: Benjamins.

Izre'el S. & Mettouchi A. (2015). Representation of speech in CorpAfroAs. Transcriptional strategies and prosodic units. In A. Mettouchi, M. Vanhove & D. Caubet (eds), *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*. Amsterdam: Benjamins: 13–41.

Martin Ph. (2003). Winpitch corpus, a software tool for alignment and analysis of large corpora. Paris: University Paris 7 Denis Diderot. https://www.researchgate.net/publication/228702364

Martin Ph. (2015). *The structure of spoken language. Intonation in romance*. Cambridge: CUP.

Moneglia M., Raso T., Mittimann M. & Mello H. (2010) Challenging the Perceptual Relevance of Prosodic Breaks in Multilingual Spontaneous Speech Corpora: C-ORAL-BRASIL / C-ORAL-ROM. In *Speech Prosody Satellite Workshop: Prosodic Prominence Perceptual and Automatic Identification*. Chicago. Université de Neuchâtel.

Moneglia M. & Raso T. (2014). Notes on the Language into Act Theory. In T. Raso & H. Mello (eds), *Spoken corpora and linguistics studies.* Amsterdam: Benjamins: 468-494.

Nicolas Martinez C. (2012). *Cor-DiAL,* Madrid: Liceus.

Nicolas Martinez C. & Lombán M. (2018). The Spanish spontaneous speech minicorpus. Architecture and comparability. *CHIMERA*, 5(2).

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

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

Raso T., Mittimann M. (2009). Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem*, 17: 73-91.

Barbosa P., Raso T. (2018) Spontaneous Speech Segmentation: Functional and Prosodic Aspects with Applications for Automatic Segmentation A, *Revista de Estudos da Linguagem*, 26(4): 1361-1396.