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


