DB-IPIC: An XML Database for Information Patterning Analysis

Alessandro PANUNZI (Università di Firenze)
Lorenzo GREGORI (Università di Firenze)

The poster will show the IPIC XML Database, which aims to represent spontaneous spoken language transcripts with three levels of annotation: prosodic boundaries, information structure and morphosyntactic tagging (PoS).

The theoretical model on which the database relies (see Cresti 2000, Moneglia and Cresti 2006) foresees that the prosody marks the referring units for the analysis of spoken language at two levels: a) the "macro"-prosody phenomena identify illocutive units; b) the "micro"-prosody phenomena identify information-structure units.

Following this model, the prosodic annotation constitutes the first level of annotation, and it is done by means of perception of tonal breaks within the speech flow: terminal breaks mark prosodically Terminated Sequences (TS), while non terminal breaks mark Tone Units (TU) within Terminated Sequences.

The prosodically TS mostly correspond to the performing of a single speech act: this is the case of the Utterances, considered as the main referring units for the analysis of the spoken language (Language into Act Theory, Cresti 2000). The other referring unit represented in the database is the Stanza (Cresti 2009), which is a terminated sequence that doesn't correspond to only one speech act (as Utterances do), but to an entire linguistic "activity". The primary intention of a Stanza is the performance of an oral text, and it develops the presentation of a thought "in process".

Within a TS, non-terminal breaks can be present, structuring the utterance into a sequence of TU which constitute a prosodic pattern (see Cresti and Moneglia 2010).

Since the Informational Patterning Theory foresees a systematic correspondence between the prosodic pattern and the information pattern of the TS, the second level of annotation provides each tonal unit with a tag regarding its information value, according to the following tagset:

a) Textual units
COM: Comment
CMM: Multiple Comment
COB: Bound Comment
TOP: Topic
TPL: Topic List
PAR: Parenthesis
APC: Appendix of Comment
APT: Appendix of Topic
INT: Locutive Introducer

b) Dialogic units
ALL: Allocutive
CNT: Conative
DCT: Dialogic Connector
EXP: Expressive
INP: Incipit
PHA: Phatic

c) Other tags
EMP: Interrupted Unit
SCA: Scanning Unit
TMT: Time taking
UNC: Unclassified

Textual IUs participate to the construction of the semantic content of the TS, while dialogic IUs are devoted to the successful pragmatic performance of the TS (e.g. to regulate the relationship between speakers, to keep open the channel of communication etc.)

This annotation has been produced using the WinPitch alignment software interface, by means of the prosodic analysis of the sound data. On the contrary, the third level of annotation, i.e. the PoS tagging, has been processed automatically, exploiting the TreeTagger software developed by the Institute for Computational Linguistics of the University of Stuttgart.

The whole annotation has been automatically converted in XML format and projected on a database. The resource runs on the eXist engine, an open source database management system that stores data according to the XML data model and features index-based XPath/XQuery processing.

A sample of an XML document containing all the annotation levels for a single turn, containing one utterance and two prosodic/information units, follows:

```
<turn speak="EDO">
<term_seq num="1" type="utt" proj_ill="unknown">
<tone_unit inf="COM" ill="none">
<word lemma="guardare" pos="VER:fin">guarda</word>
<word lemma="chi" pos="WH">chi</word>
<word lemma="c'" pos="ADV">c'</word>
<word lemma="essere" pos="VER:fin">è</word>
<break type="nonterminal">/</break>
</tone_unit>
<tone_unit inf="ALL">
<word lemma="nonna" pos="NOUN">nonna</word>
<break type="terminal">//</break>
</tone_unit>
</term_seq>
</turn>
```

The database currently contains data from the informal section of the C-ORAL-ROM Italian Corpus (see Cresti and Moneglia 2005). This collection is constituted by 74 total texts, which correspond to more than 21000 TS and roughly 125000 words.

Moreover, a Brazilian Portuguese minicorpus from the C-ORAL-BRASIL Corpus (see has been implemented (roughly 30000 words in more than 5500 TS).

The architecture of the database allows cross-level queries on the data. For this purpose, a specific interface has been developed (to be shown as a demo-software during the poster session). Quantitative data about the information patterning of Utterances and Stanzas and the morphosyntactic and lexical fillings of different IUs will be also presented.

References

Cresti, E. 2000. Corpus di italiano parlato, 2 voll., CD-ROM. Firenze: Accademia della Crusca.

Cresti, E. and M. Moneglia (eds). 2005. C-ORAL-ROM. Integrated reference corpora for spoken romance languages, DVD + vol. Amsterdam: Benjamins.

Cresti, E. and M. Moneglia 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: FUP.

eXist. http://exist.sourceforge.net/

Moneglia, M. and E. Cresti. 2006. C-ORAL-ROM. Prosodic boundaries for spontaneous speech analysis. In Y. Kawaguchi, S. Zaima and T. Takagaki (eds), Spoken Language Corpus and Linguistics Informatics. Amsterdam: Benjamins, 89-114.

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.

Treetagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

WinPitch. http://www.winpitch.com/

XML. http://www.w3.org/XML/

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