

C-ORAL-ROM - Integrated Reference Corpora for Spoken Romance Languages. Corpus and Research

Emanuela Cresti (LABLITA, University of Florence)

The C-ORAL-ROM multilingual resource provides a comparable set of corpora of spontaneous spoken language of the main romance languages, namely French, Italian, Portuguese and Spanish. The resource is the result of the C-ORAL-ROM project, which has been undertaken by an European consortium co-ordinated by the University of Florence, (University of Florence, Université de Provence; Centro de Linguística da Universidade de Lisboa (CLUL); Universidad Autónoma de *Madrid*) and funded within the 5th EU framework program.

C-ORAL-ROM consists of 772 spoken texts and 123:27:35 hours of speech. Four comparable recording collections of Italian, French, Portuguese and Spanish spontaneous speech sessions (roughly 300,000 words for each Language)

Each recorded session of the corpus is associate to the following main annotations:

- a. Session metadata (in CHAT and IMDI format)
- b. The orthographic transcription,(in CHAT format; Mac Whinney, 1994) (<http://childes.psy.cmu.edu/manuals/CHAT.pdf>), enriched by the tagging of terminal and non terminal prosodic breaks and simultaneously (in txt files)
- c. The text-to-speech synchronization, based on the alignment to the acoustic source of each transcribed utterance, in .xml files.
- d. Textual resource with Part of Speech (PoS) and lemma tagging of each form

The collection of files is conceived to allow maximum and easy exploitation of the linguistic information recorded in the corpus. Special attention is devoted to the annotation of the reference units of analysis for spontaneous speech (utterance) and to the synchronization of the transcripts to the acoustic source.

C-ORAL-ROM is available in two multimedia editions:

1. Through the ELDA Catalogue in 9 DVDs where files are non-compressed and non-encrypted. This form is devoted to speech laboratories. (<http://www.elda.org/catalogue/en/speech/S0172.html>)
2. Through Benjamins Publishing company, which presents the resource in compressed and encrypted format in only one DVD accompanied by an explicative book. This form, that does not allow copying, is designed for wide distribution in the linguistic community.

Main topics of the presentation

The main C-ORAL-ROM objective is to allow Linguistic Studies and Natural Language technologies to face challenging LRs which represent spontaneous speech in real environment. To this end C-ORAL-ROM “aims to represent the variety of speech acts performed in everyday language and to enable the induction of prosodic and syntactic structures in the four romance languages, from a quantitative and qualitative point of view”. This task is very ambitious and requires an adequate representation of both the spoken universe (Corpus design), and a linguistic annotation that is also adequate to speech.

The presentation will focus on the corpus design structure of the C-ORAL-ROM resource and will briefly present the annotation strategy chosen in C-ORAL-ROM for what concerns the main unit of analysis of spontaneous speech. The validity of the C-ORAL-ROM assumptions at both levels will be supported showing the generalizations that have been obtained through the early cross-linguistic studies accomplished on the C-ORAL-ROM corpus.

The corpus design structure of the resource and the parameters chosen to allow the comparability of the four language will be presented making reference to other large spoken corpora initiatives (CoSIH, Dutch Corpus). According to the tradition of socio-linguistic studies (see Berruto 1987; Biber 1988; Biber et al. 1998; Gadet 1996, De Mauro et Al. 1993) that highlight the significance of the sociological and contextual parameters in the definition of speech qualities C-ORAL-ROM records spoken romance languages in a huge variety of contexts. Its design (like the Dutch Corpus) is primarily based on contextual traits (diaphasic variation) and adopts a set of new solutions for the representation of spontaneous speech in the formal and informal contexts.

The unit of reference for spontaneous speech performance is defined as the utterance (Cresti 2000 - Quirk et al 1985); i.e. the linguistic entity which copes with a speech act (Austin 1962). The utterance can be identified in speech by its terminal prosodic break. All the romance corpus has been tagged by terminal prosodic breaks and all the utterances, so devised, have been aligned with their acoustic counterpart.

The definition of the utterance boundaries through its prosodic properties allows a proper analysis of the four spoken romance corpora and permit to carry out their general comparison from a linguistic point of view (Cresti & Moneglia (2005). The validity of the corpus design and the relevance of corpus annotation will be tested presenting the results of some comparative studies accomplished on the four romance languages.

Three kinds of basic cross-linguistic measurements of speech show regular and sensible context-bound variations. Such measurements regard: a) the distribution of Part of Speech in the speech performance; b) the weight of the utterance, in terms of length and speed and their correlations with the weight of the dialogic turn; c) the main structural strategies used by speakers to build up the utterance in spoken language.

The presentation will show that in all language corpora such general linguistic properties co-vary in connection with the contextual variations of the corpus design. This leads to the conclusion that crucial spoken language behaviors are required by contextual features and that therefore the representation of spoken language activity calls for an adequate representation of contextual variation, as proposed by C-ORAL-ROM and other large spoken corpora initiatives in order to capture relevant qualities of language use.

References

- Austin, L.J., 1962. *How to Do Things with Words*. Oxford: Oxford University Press.
- Berruto, G. 1987. *Sociolinguistica dell'Italiano Contemporaneo*. Roma: La Nuova Italia Scientifica.
- Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press
- British National Corpus <http://www.natcorp.ox.ac.uk/>
- CoSIH CORPUS <http://www.tau.ac.il/humanities/semitic/cosih.html>
- Cresti, E. 2000. *Corpus di Italiano Parlato*, voll. I-II, CD-ROM. Firenze: Accademia della Crusca.
- Cresti, E. & Moneglia, M. (eds.). 2005. *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.
- De Mauro, T., Mancini, F., Vedovelli, M. and Voghera, M. 1993. *Lessico di Frequenza dell'Italiano Parlato*. Milano: ETAS.
- Dutch Corpus http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm
- Gadet, F. 1996. "Variabilité, variation, variété". *Journal of French Language Studies* 1: 75-98.
- Halliday, M.A.K. 1989. *Spoken and Written Languages*. Oxford: Oxford University Press.
- Izre'el, S. 2006. *Intonation Units and the Structure of Spontaneous Spoken Language: A view from Hebrew*. Ms. Tel-Aviv University.
- Izre'el, S., Hary, B. and Rahav, G. 2001. "Designing CoSIH: The corpus of spoken Israeli Hebrew". *International Journal of Corpus Linguistics* 6: 171-197.
- Leech, G., Rayson, P. & Wilson, A, 2001. *Word Frequencies in Written and Spoken English*, London: Longman.
- Karcevsky, S. 1931. "Sur la phonologie de la phrase". *Travaux du Cercle Linguistique de Prague* IV: 188-228.
- MacWhinney, B. 1994. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Massimo, M. (2004). "Measurements of Spoken Language Variability in a Multilingual Corpus. Predictable Aspects" in M.T Lino, M.F. Xavier, F. Ferraira, R. Costa, R. Silva (eds) *Proceeding of the 4th LREC Conference*, ELRA, Paris, vol 4 pp 1419-1422
- Moneglia M. (2004b) *Specifications of the C-ORAL-ROM corpus*. ELRA, Paris
- Moneglia, M. (to appear) Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective. To appear in Kawaguchi et al. *Usage Based Linguistic Informatics* 4, Benjamins
- Quirk, R. S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.