C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages

PARTNERSHIP:
- Università degli studi di Firenze (UFIR.DIT) - Italy - Coordinator
- Fundação da Universidade de Lisboa - Centro de Linguística da Universidade de Lisboa (FUL-CLUL) – Portugal
  Maria Fernanda Bacelar do Nascimento, José Bettencourt Gonçalves, Rita Veloso, Sandra Antunes, Nuno Martins, Florbela Barreto, Raquel Amaro
- Université de Provence (UPRO) - France
- Pitch Instruments France S.A.R.L. (PITCHFRANCE)
- European Language Resources Distribution Agency S.A.R.L. (ELDA) - Instituto Cervantes (IC)

PROJECT OVERVIEW
The C-ORAL-ROM resource is a multilingual corpus of spoken language for the main romance languages, namely Spanish, Portuguese, French and Italian, constituted by formal and informal speech, in a total of 1,200,000 words (300,000 words for each language).

The project involved the following tasks:
- orthographic transcription, in chat format, enriched with the tagging of terminal and non-terminal prosodic breaks and session metadata;
- text-to-sound synchronization, in WinPitch Corpus format, based on the alignment of each transcribed utterance;
- lemmatization and PoS tagging.

This resource comprises several components:
- a multimedia corpus;
- software tools for speech analysis;
- concordances extraction tool.

C-ORAL-ROM is available in two formats:
1. 8 DVD’s with full access to explore the materials;
2. an encrypted version, available in 1 DVD, which accompanies the book published by John Benjamins Publishing Company (2005), containing comparative linguistic studies and standard linguistic measures of spoken language variability derived from corpora analysis.

Funding: Information Society Technologies Programme (IST-2000-26228) – European Commission