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)**
- **Universidade Autónoma de Madrid (UAM) – Spain**
- **Instituto Trentino di Cultura (ITC-irst)**
- **Instituto Cervantes (IC)**

http://lablita.dit.unifit.it/coralrom/
http://www.clul.ul.pt/english/sectores/projecto_coralrom.html

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:

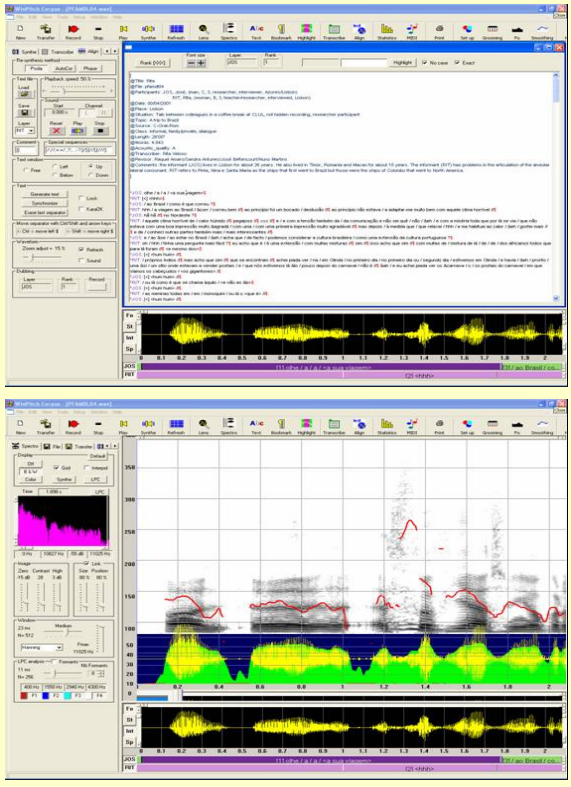
1. 8 DVD's with full access to explore the materials;
2. an encrypted version, available in 1 DVD, which accompanies the book published by John Benjamins Publishing Company (2005), containing comparative linguistic studies and standard linguistic measures of spoken language variability derived from corpora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

PORTUGUESE CORPUS CONSTITUTION (318,593w)

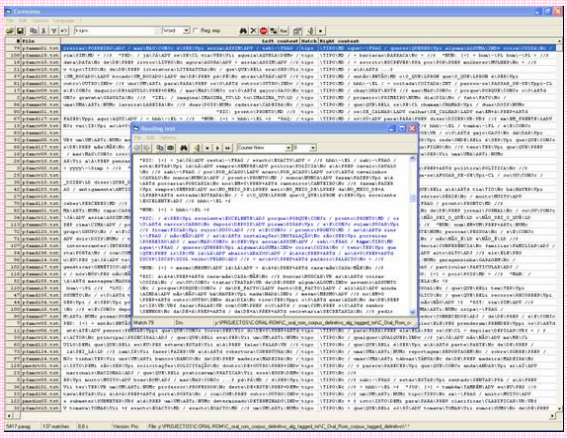
INFORMAL			
Family /	Conversations	24.449	
/ Private	Dialogs	62.738	
	Monologs	46.005	**133.192**
Public	Conversations	1.817	
	Dialogs	23.119	
	Monologs	7.710	**32.646**
TOTAL			**165.838**

FORMAL			
Natural	Business	10.215	
Context	Conferences	9.750	
	Law	6.315	
	Political Debate	8.923	
	Prof. Explanation	6.473	
	Preaching	6.127	
	Political Speech	8.649	
	Teaching	9.822	**66.274**
Media	Interviews	14.570	
	Meteo	1.930	
	Reportages	10.762	
	Scientific Press	9.923	
	Sport	5.676	
	Talk Show	17.396	
	News	1.859	**62.116**
Telephone	Private		**24.365**
TOTAL			**152.755**

WINPITCH CORPUS – SPEECH SOFTWARE



CONTEXTES – CONCORDANCES EXTRACTION TOOL



C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages
E. Cresti and M. Moneglia (eds.)
John Benjamins Publishing Company, 2005

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