C-ORAL-ROM
THE SPOKEN ROMANCE CORPUS: COMPARABILITY IN A MULTILINGAL GENERAL RESOURCE OF SPONTANEOUS SPEECH

M.Moneglia, E.Cresti
Dept. of Italian, University of Florence

Language technology in a multilingual society needs huge language data bases, especially for validation of technologies based on spoken language interface. However such resources are frequently difficult to be reused especially for comparison , because of their different formats, constitution criteria, and degree of accessibility.
The paper will present a project which is intended to provide the linguistic community and the TAL community with a comparable set of corpora of spoken language for the main romance languages, namely French, Italian, Portuguese and Spanish, where textual information and sound source will be associated in a multimedia edition. The resulting multimedia spoken romance corpus, published on DVDs , will be integrated with: a) high performance tools for the analysis of both sound and text; b) studies on paper offering standard linguistic measures for spoken romance languages derived from corpora analysis. Generally speaking, the project will provide an easy access LR suitable for spoken language comparison in the romance area and extremely useful for validation of both speech and textual multilingual HLT.
The background of the project can be found in particular in : a) parallel editions of restricted sampling of each corpus (*Corpus dell'italiano parlato: campioni*, Florence, Accademia della Crusca, in press; *Corpus du Français parlé*, INaLF Paris, in press; *Corpus de referencia do portugues contemporaneo*, CLUL, Lisbon; *Corpus Oral Peninsular*, Universidad Autonoma, Madrid); b) recent international meetings on corpus linguistics (*Questions de methode dans la linguistique sur corpus*, Perpignan, May 1998; *Macrosyntax and pragmatics, the linguistic analysis of spoken language*, Florence, April 1999).
The paper will focus on the constitution criteria which are needed in order to ensure comparability in the multilingual resource and validation of HLT on a general resource of spontaneous spoken language.
The spoken romance corpus will be extracted from already existing language resources, (as a significant sampling, roughly 400.000 words for each language), that each partner has already set up. Crucially, sampling operation will be performed applying to each corpus a strictly defined set of criteria which will ensure documentation of variation across language uses. (Cfr. Bilger 1997; Labov, 1966; Biber 1994; 1988; Berruto, 1987; Gadet; 1996; Bacelar do Nascimento,1998; Moneglia 1999). The romance corpus will testify for each language: a) formal speech; b) informal speech; c) media production. In particular each corpus will cover a huge proportion of spontaneous/informal speech (50% of the total), offering the base for a better documentation of its general peculiar properties.
All texts will be revised from their original format and will be proposed in the same format and with the same degree of transcription accuracy. The transcription format will follow European and de facto standard (Gibbon et alii, 1997; Mac Whinney, 1994) and will be systematically annoted with respect to prosodic parsing ('t Hart et alii, 1990; Cresti, 1994)
The resulting *Spoken romance corpus,* realized in a multimedia edition on DVDs, will be integrated with tools allowing direct access to both concordances of the text and analysis of the acoustic signal (WINPITCHCORPUS. Cfr http://winpitch.com). WINPITCHCORPUS will be the result of the implementation, for application on large corpora, of a high quality language tool already conceived in the EC (WINPITCH), which offers real-time analysis of the signal with respect to

main vocal parameters: Fo, duration, intensity, spectrum and a set of easy synthesis options, extremely interesting for testing and validation procedures. All texts will be offered in a naked ASCII format and in SGML format and will be aligned to their sound source which will be given in uncompressed standard format (22050 Hz 16 bit in .wav format).

The choice of standard formats for both text and sound file will allow access to most common tools for textual and vocal analysis and will ensure maximum results in the exploitation and reusability of the resource. In order to ensure better validation procedures on a general resource, standard linguistic measures of spoken language will be provided. In particular a series of quantitative measures will be given: density, words/fragments proportion; middle utterance length; tone segmentation average; middle length of dialogical turns; etc. A gender classification of spoken texts will be proposed on the basis of their internal linguistic properties.

The simultaneous text-signal analysis on a spontaneous spoken LR in a multimedia edition is one of the main quality of the romance corpus. Multimedia output enables the direct use of the same LR in order to validate both kind of instruments devoted to voice technology: automatic voice recognition (sound to text interface) and vocal syntesis (text to sound interface). The corpus will allow direct comparison of syntetic with natural signals, realized in varieties of spontaneous production conditions, and testing correct recognition .