Integrated tools for exploitation of a spontaneous speech corpus of Spanish

Antonio MORENO-SANDOVAL and José M. GUIRAO

Abstract

This paper presents a query system to a corpus of spontaneous speech, concretely the Spanish C-ORAL-ROM corpus (Cresti & Moneglia eds. 2005; Moreno et al. 2005). The corpus consists of 181 transcripted sessions, in different registers and communicative situations. With over 42 hours of recorded data and almost 500 speakers, the corpus has 312,000 tokens (words) of 21,000 different types (lemmas).

The system will be accessed through a web page, although it will eventually be also available as an independent application. The tools consist of three main components:

1. A metadata query system, which retrieves texts with features specified in the query. Those features include sociolinguistic information about the speakers, as well as contextual and thematic information about the recording.

2. A concordancer of text and sound: the system looks up words or multi-words expressions in all the texts and retrieves every "utterance" where the searched string appears along with the original sound fragment.

3. A morphological processor of Spanish, based on broad-coverage lexicon, which provides all the possible analyses for a given wordform. It can be used as a Part-of-Speech tagger for sentences in Spanish.

Finally, some examples of its application to teaching/learning Spanish as a second language, as well as to describing properties of spoken Spanish will be shown.