Relating Lexical Items to Sociolinguistic Features in a Spontaneous Speech Corpus of Spanish

José María Guirao Miras
Departamento de Lenguajes y Sistemas Informáticos, Escuela Técnica Superior de Ingeniería Informática, Universidad de Granada
c/ Daniel Saucedo Aranda s/n, 18071 Granada, Spain
jmguirao@ugr.es

Ana González Ledesma
Guillermo de la Madrid Heitzmann
Manuel Alcántara Plá
Antonio Moreno Sandoval
Laboratorio de Lingüística Informática, Departamento de Lingüística, Universidad Autónoma de Madrid
Carretera de Colmenar Viejo Km.15, Cantoblanco, 28049 Madrid, Spain
ana.ledesma@adi.uam.es, guille@maria.lllf.uam.es, manuel@maria.lllf.uam.es, sandoval@maria.lllf.uam.es

This paper shows the application of statistical tests to a spontaneous speech corpus of Spanish. Our goal is to find representative differences between different parts of the corpus. To this end, we tagged n-grams in the corpus with features related to the speaker (age, gender, etc), or the context (dialogue, monologue, media, etc), and applied the log-likelihood test (Dunning 1993) in order to find the most representative lexical items for each specific feature.

The paper is divided in two sections. In the first part, the characteristics of the spoken corpus are shown together with the explanation of the computational tool. In the second part, a first rough estimate of the results obtained is given, as well as possible applications of the model.

We work with the Spanish part of the multilingual corpus C-ORAL-ROM, which is made up of 300,000 words. These have been orthographically transcribed. The domain distribution is diversified in order to allow cross-linguistic comparison. The corpus is organized using three variables: situation (formal/informal), channel (media, telephone, natural context) and number of speakers (dialogue, monologue, conversation). Each text contains the transcription and a header where the information about the gender, age, education, occupation and geographical origin of the speakers is shown.

We have developed a tool that permits the clustering of those words or multi-words that appear three or more times and sorts them according to their frequency of use. The corpus is XML-tagged, which allows to automatically relate these results to the sociolinguistic variables (age, gender, etc) and to the text typology (formal, informal, media...). The access to the data is possible through two ways. The first one provides the sorted list of words and clusters, and from each item one can reach the corresponding sociolinguistic and text-typologic information. The second one works the other way around: from a list of sociolinguistic and typological items one obtains the sorted list of words and clusters for that field (men, formal, media...).

The second part of the paper is devoted to the analysis and explanation of these results. Here we show how the frequency of words and clusters varies noticeably according to the different parameters. For instance, the discourse marker "o sea" (*that is*) is much more frequent in informal contexts, and the frequency of use of the expression "no sé" (*I don't know*) in feminine speakers is double for male speakers.

This paper ends by stating the importance of this procedure as an empirical method for the validation of sociolinguistic hypotheses in spoken language.