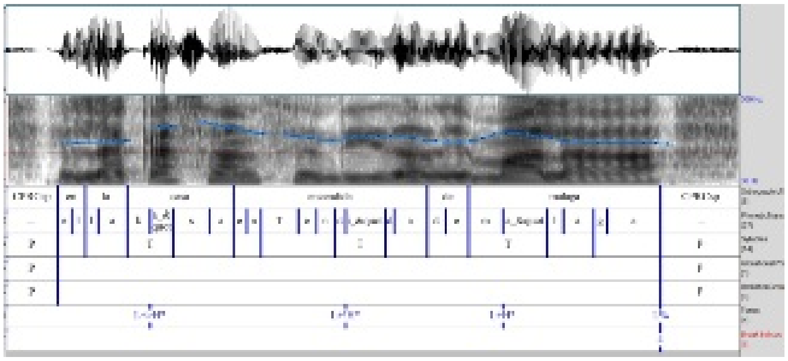
UABDIVULGA
BARCELONA RECERCA I INNOVACIÓ

05/2014

Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan



A literature review on prosody studies reveals a lack of Catalan and Spanish corpora. Glissando is a Spanish and Catalan corpus which intends not only to fill this gap but furthermore show the capabilities of extensive speech corpora for the empirical study of prosody. With the aid of this corpus, it will be possible to analyse differences due to linguistic and sociolinguistic criteria such as genre, speech style, and voice register. Within this domain, the corpus has been proven to be a useful resource for teaching tasks related to linguistics, oral expression, and communication.

The Glissando corpus has been developed within the framework of “Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan” (FFI2008-04982-C03-02/FILO), a research project coordinated between the University of Valladolid (ECA-SIMM group, D. Escudero), the Pompeu Fabra University (Department of Linguistics and Communication, J.M. Garrido), and the Universitat Autònoma de Barcelona (Department of Hispanic Philology, Lourdes Aguilar). The corpus comprises two distinct data-sets, a news subcorpus and a dialogue subcorpus, the latter containing either unplanned conversations or task dialogues oriented to a specific goal in the domain of information request (travel information, information request for an exchange university course, information request for a tourist route, etc.).

The recordings were made in high acoustic quality by two profiles of speakers: broadcasting and advertising professionals, and native undergraduate students. The twenty-five hours of recordings cover different reading styles (radio, advertising and neutral), registers (reading news, formal and informal dialogue), voices (male and female), and languages (central Catalan and standard European Spanish). The inclusion of these variables aims to facilitate cross-linguistic, interspeaker, and inter-style analyses.

The entire material of the Glissando corpus has been orthographically and phonetically transcribed, aligned with the acoustic signal, and prosodically annotated. The image here included exemplifies what can be obtained with a computer programme for the analysis of speech (i.e. Praat). Several types of information are included in the image: information pertaining to lexical words, syllabic borders, lexical accents placement (identified by "_&quot"), tonal prominences (symbolized by T), and the silences present in the speech chain (shown with the letter P, i.e. Pause). From this information, users can practice the phonetic transcription of the sounds, isolating those that show the high degree of difficulty in discrimination. Also, they may compare stress levels in syllables (e.g. primary versus secondary accents) or observe to what degree the duration of pauses affects the understanding of the message.

In order to represent intonation patterns, we have applied the ToBI intonational phonology model developed in previous works for Spanish and Catalan (Sp_ToBI http://prosodia.upf.edu/sp_tobi/en/and Cat_ToBI http://prosodia.upf.edu/cat_tobi/en/). This is a complete system of intonational representation that allows us to describe the melodic movements taking into account the difference between low tones (i.e. L) and high tones (i.e. H) as well as to distinguish a hierarchy of prosodic domains: intonational phrases that involve complete meaning (i.e. boundaries characterized with number 4) and intermediate phrases (i.e. boundaries characterized with number 3). The corpus has been designed to cover the research needs of the groups involved in the project, but its possibilities for research are numerous. The audio files with their associated tags are freely available for research purposes in the web page of the project: http://veus.glicom.upf.edu

The most direct use of Glissando corpus is closely related to university teaching and research, but applications to high school teaching are also possible, considering how common it is to employ new technologies for such purposes. Indeed, for these students gaining access to speech corpora is a way to improve their communicative skills, since it allows them to train as good speakers or simply to make oral practices in order to distinguish between colloquial and formal intonation. The main aim is to provide tools which enable them to incorporate different speaking styles and discursive practices in their daily life.

Finally, in addition to the descriptive and theoretical research that can be developed through the Glissando corpus, the prosodic annotation of corpus has proven to be useful, for example, for the automatic prosodic description and modelling of intonation, including the development of tools for Automatic Speech Recognition or Dialogue Systems. In particular, an annotated sample of Glissando corpus has served to create a semi-automatic tool for prosodic transcription and nowadays we are evaluating its utility as a support material for the creation of production and/or comprehension tools and the learning of Spanish and Catalan as a foreign language.

Yurena María Gutiérrez
Lourdes Aguilar
yurenagg@gmail.com, Lourdes.Aguilar@uab.cat

References

Garrido, J. M.; Escudero, D.; Aguilar, L.; Cardeñoso, V.; Rodero, E.; De-La-Mota, C.; González, C.; Rustullet, S.; Larrea, O.; Laplaza, Y.; Vizcaíno, F.; Cabrera, M., Bonafonte, A. Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. Language Resources and Evaluation 47(4): 945-971. 2013.

View low-bandwidth version