

CORPORA OF LATIN AMERICAN SPANISH FOR RESEARCH IN PROSODY AND SYNTHESIS

Alejandro C. Renato

Genius Institute of Technology. Manaus. Brazil

José A. Alvarez

Faculty of Exact and Natural Sciences.UBA. Argentina

ABSTRACT

The present article describes the creation, labelling and main characteristics of a corpus of spoken Latin American Spanish. The corpus was collected with several objectives in mind: a) to fulfill our own research needs in the study of Latin American Spanish prosodic phenomena, where the absence of available corpora has already been noticed [1][6], b) to be able to experiment with prosodic models used in speech synthesis and c) to make it available to the community of researchers.

1. INTRODUCTION

Internet radio corpora have been used mainly in spoken news recognition. However, we found that many samples have an acceptable quality for other types of research: they are recorded in audio studios in 16 bits, more than 16000 Hertz of sampling rate, a relation signal-noise greater than 30 db and spoken by professional speakers. The audio format “mp3” was found to have better quality, while others damage the main characteristics of waveforms.

Because many spanish speaking countries have very limited resources for research in these subjects, an important advantage of this kind of material is economical. In Latin America, there are 19 Spanish speaking countries with several dialects in each of them, so that a corpus representative of this wide diversity is very difficult to achieve, specially considering that only a few dialects have a commercial interest. From a methodological point of view, a corpus with this characteristics is an important source for hypothesis testing and a valuable resource for the development of more precise tools and a higher degree of automatization to obtain annotations in corpora, for example, for pitch events, prominence, etc [3].

The methodology usually followed by prosodic studies is to produce models on a reduced data set and later generalize them. The method we used is rather different: it consists in testing hypotheses in a large

corpora, and later, make experiments in the laboratory on particular aspects.

With respect to synthesis, we found that the needs for the selection of units are different for segments and for prosodical features: whereas for segments in Spanish a set of about 400 to 1200 sentences may be required, depending on the domain in question; in prosodic training the number of sentences is more difficult to determine, because we do not have yet objective methods to measure the quality of prosodical synthesis. Having two different corpora from each speaker for different necessities, could prove to be a provisional solution to this problem [10]. We have also found the need to have a database as a testbed of the robustness of prosodic algorithms and methods with respect to dialects and idiosyncratic characteristics of speakers and styles.

It should be noticed, however, that for synthesis, the corpus has the disadvantage of sentences taken from different days of broadcasting of a program, and the lack of EGG parameters.

2. CORPORA DESCRIPTION

The sources of the corpus are international radio broadcastings transmitted by Internet in Spanish of different countries of Latin America. The collection of waveforms was started in 1998 and continues to the present time.

The units used for segmentation are sentences, when the text is read, and utterances with unitary sense for spontaneous speech, divided by sustantive pauses. Segmentation was done using the Transcriber [2] software in combination with ch_wave, part of the Speech Tools package [8]. Utterances having superpimposed speech, music or background noise, non isolatable noises from the speaker or low quality of audio, are discarded, unless they posses a dialectal value. The corpus is divided in 5 separate subsets: 1) 1200 sentences by eight newscasters, 2) interviews or conferences by personalities of Latin American culture of 30 to 60 minutes of duration (ten speakers), 3) political speeches

before the Assembly of the UN of seven latin-american presidents, 4) reading of literature and letters, and 5) more than 80 speakers in spontaneous speech during 4 to 10 minutes and belonging to different regions from Latin America.

3. TRANSCRIPTION

Segmental transcription was done by automatic forced alignment with speaker dependent models. For the segmental transcription we used the SAMPA [5] alphabet for Spanish; syllables and phone segmentation was corrected by hand. Prosodic annotation is made on separate files for syllable prominence and breaks. For tonal transcription we used the extended Tilt annotation [9], which we found adequate for Spanish and from which other kinds of annotations, such as ToBI, may be derived.

4. SELECTION OF SENTENCES

The selection of sentences is done from the corpus of radial news. For this purpose, the selected sentences are transcribed into phones and syllables, and a balancing is done using the CorpusCrt program [7]. In this way we extract a corpus of sentences balanced by phones, syllables and demissyllables. With a phonetic transcription made on the base of 35 phones, including accentuated vowels, we selected 1200 sentences for each of four speaker from an average of 2500 sentences for each - 10000 altogether -, and 180.000 words, with the following characteristics:

	A	B	C
Phones	+140/phones	-	-
Demissyll	727	876	1747
Syllables	1789	2378	4810

Figure 1. A) Average of each corpus. B) Average of the whole corpus. C) Reference Dictionary

The first column is the average of units for each corpus. The second column refers to the whole corpus (4 speakers) and the third column indicates the amount of units found in the reference dictionary used, which contains 110.000 entries. Of the 4810 different syllables found, 3352 have a single occurrence, and 1458 have a frequency of 2 or more. All of the last ones are included in the corpora. We find an analogous situation with demissyllables. The phone with the least occurrence is \g \ with a frequency of 140, whereas the one of greatest occurrence is \e \ with a frequency of 20211. We tried to increase the number of sentences without obtaining greater quantities of different units.

5. STATE OF ADVANCEMENT

The degree of advancement of the project is 60 % for a total of 30 hours collected. A part-of-speech tagging of the corpora was made, but still has to be corrected. In addition, a parser using an HPSG grammar will be implemented for syntactic labelling.

Because one of the objectives is to make the corpus available through Internet, inquiries to the radial broadcasters are being done about copyrights. The feasibility to make the corpora available arises from previous experiences [4], where the mass media companies and publishing houses released part of the rights on the material for research purposes.

6. REFERENCES

- [1] Beckman, Mary, Díaz-Campos, Manuel, McGory, Julia, Morgan, Terrell. "Intonation across Spanish, in the Tones and Break Indices framework". University of Ohio. Linguistics Department.1999.
- [2] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech". First International Conference on Language Resources and Evaluation (LREC) pages 1373-1376.1998.
- [3] Huang Xuedong, Acero, Alex, Hon, Hsiao-Wuen. "Spoken Language Processing: A Guide to Theory, Algorithm and System Development". Prentice Hall. 2001.
- [4] Marcos Marín, Francisco. "Corpus Lingüístico de Referencia del Español: Argentina y Chile". Universidad Autónoma de Madrid. 1992.
- [5] Mariño, José, Moreno, Asunción. "European and Latin American Spanish Allophone set". Project SALA. In: <http://www.sala2.org/SALASAMPA.rtf>
- [6] McGory, Julia, Díaz-Campos, Manuel. "SpanishToBI Workshop". In <http://www.ling.ohio-state.edu/~tobi/sp-tobi/spanish.html>. 1999.
- [7] Sesma Bailador, Alberto. "CorpusCrt". Politechnic University of Catalonia.1998.
- [8] Taylor, Paul, Caley, Richard, Black, Alan and King, Simon. "The Edinburgh Speech Tools Library". In http://www.cstr.ed.ac.uk/projects/speech_tools/.1997.
- [9] Taylor, Paul."Analysis and Synthesis of Intonation using the Tilt Model". JASA, vol 107 3, pp. 1697-1714.Year 2000.
- [10] Raux, A, and Black, A. "A Unit Selection Approach to F0 Modeling and Its Application to Emphasis". ASRU 2003, St Thomas, US Virgin.