Experiments on the Construction of a Phonetically Balanced Corpus from the Web

L. Villaseñor-Pineda ‡, **M. Montes-y-Gómez** ‡, **D. Vaufreydaz** * and **J-F. Serignat** *

‡ Laboratorio de Tecnologías del Lenguaje, INAOE, México
{villasen, mmontesg}@inaoep.mx
* Laboratoire CLIPS/IMAG, France
{Dominique.Vaufreydaz, Jean-Francois.Serignat}@imag.fr

Abstract. The construction of a speech recognition system requires a recorded set of phrases to compute the pertinent acoustic models. This set of phrases must be phonetically rich and balanced in order to obtain a robust recognizer. By tradition, this set is defined manually implicating a great human effort. In this paper we propose an automated method for assembling a phonetically balanced corpus (set of phrases) from the Web. The proposed method was used to construct a phonetically balanced corpus for the Mexican Spanish language.

1 Introduction

The construction of a speech recognition system requires a set of recordings to obtain the pertinent acoustic models. These recordings must consider several aspects in order to produce a robust recognizer. For instance, (i) the spoken corpus must be *rich*, i.e., it must contain all the phonemes of the language, and (ii) it must be *balanced*, i.e., it must preserve the phonetic distribution of the language.

The construction of a phonetically rich and balanced corpus is based on the selection of a set of *phrases* that will be recorded. Traditionally, this selection involves a great human effort. First, it is necessary to select a set of words phonetically rich, and join them to form the desired phrases. Later on, it is necessary to verify the phonetic distribution of the constructed phrases, and if required, add and delete some phrases. Certainly, these changes affect the overall phonetic distribution, and thus, the process must be repeated until an adequate distribution is reached.

In this paper, we propose a straightforward method for selecting a set of phrases to be recorded. This method is entirely different from the traditional process. It is supported on the hypothesis that the Web, for its huge size, is already a phonetically rich and balanced source, and thus, taking a subset of it is enough to assemble a phonetically rich and balanced corpus.

The following sections describe the proposed method, and illustrate the construction of a phonetically rich and balanced corpus for the *Mexican Spanish language*.

2 Collecting documents from the Web

In order to assemble the desired corpus, we first need to collect a set of documents from the web (a broad exposition on this problem was presented in [2, 3]). For this

purpose, we used the CLIPS-Index web robot [1]. This robot starts from an initial set of URLs, and gathers all their web pages (in simple HTML format) and text documents. This robot also allows filtering the web pages in accordance with a domain of interest. In our case, we downloaded only the pages from the Mexican domains.

Additionally, we deleted all the tags, headers and other metadata from the downloaded web pages and documents. After this process, we obtained a text corpus, presumably in Mexican Spanish, of 1.2 Gbytes, with a total of 244,251,605 words and 15,081,123 lines.

3 Selecting a set of phrases for recording

The text corpus collected from the web was our raw material. From this corpus we selected the phrases containing only Spanish words (a lexicon was used for this task) and having more than 30 words.

Initially, we used a lexicon of 177,290 lexical forms obtained from two Spanish dictionaries and several Mexican newspapers and magazines. Using this lexicon we selected a primary set of phrases called Corpus170.

Because we considered that the initial lexicon was not bigger enough for the task at hand, we performed another experiment with an enlarged lexicon of 235,891 lexical forms. This new lexicon was constructed as follows. First, based on the initial lexicon, we extracted the unknown words from the text corpus (refer to the section 2). Then, using Google, we looked for the Spanish web pages containing the unknown words. Finally, we counted the occurrences for each word in the returned pages, and aggregated to the initial lexicon those having an occurrence greater than some given threshold. Using this enriched lexicon we obtained a new set of phrases called Corpus230.

The table 1 shows the main characteristics of both sets of phrases, the Corpus170 and the Corpus230.

Table 1. The collections Corpus170 and Corpus230

	Lexicon Size	Number of phrases	Number of words	Number of words per phrase
Corpus170	177,290	339,833	14,511,061	42.7
Corpus230	235,891	344,619	14,766,638	42.8

3.1 Phonetic distribution

In order to evaluate the quality (i.e., richness and balance) of the selected set of phrases, we compared its phonetic distribution[1] with the phonetic distribution of the Spanish language reported in the literature. The figure 1 shows the phonetic distribution of our corpora as well as the Spanish phonetic distribution in accordance

[1] We transformed the words of the corpora to their corresponding phonemes using a tool developed in the DIME project [4].

with a Latin-american [5] and Iberian [6] studies. From this figure we get the following conclusions:

1. *Our initial hypothesis is correct*; the phonetic distribution of the corpora obtained from the web is very close to those reported for the Spanish language. For instance, the correlation coefficient between the Corpus230 and the Latin-american and Iberian studies were 0.994 and 0.942 respectively.

2. *The size of the lexicon and the corpus is not a fundamental element*. For instance, the Corpus170 presents a notable correspondence to the Corpus230 (with a correlation coefficient of 0.99), even when they were built using lexicons with more than 60 thousand different lexical forms.

Figure 1 also shows that the generated corpora are phonetically more similar to the Latin-american Spanish than to the Iberian. The occurrence proportion of the phonemes /a/ and /e/ are good examples of this circumstance.

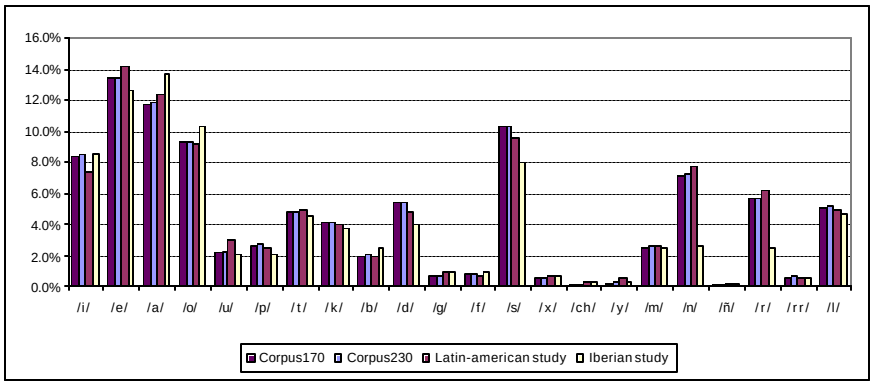

Figure 1. Phonetic distribution of the generated corpora

3.2 Corpus reduction

Approximately, only 6000 phrases are required in order to construct a 10 hours of recorded corpus. An automatic method to select the best phrases from the generated corpora consists in computing the perplexity of each phrase in accordance with a language model, and keeping those phrases having the lower perplexity. We constructed the language model from a collection of written conversations among several individuals. This collection is of 4.8Mb and has 864,166 words, and 20893 lexical forms. Using this language model we obtain a set of 6082 phrases with a perplexity less than 7.25. The table 2 shows the main characteristics of this set of phrases.

It is important to mention that the phonetic distribution of the reduced corpus maintains a strong correlation with the Latin-american Spanish phonetic distribution (0.994) and with the Corpus 230 distribution (0.997).

Table 2. Corpus reduction

	Lexicon Size	Number of Phrases	Number of words	Number of words per phrase
Reduced Corpus	235,891	6082	220,776	36.3

4 Future work

Before recording the corpus, it will be necessary to manually ensure its content. Basically, we plan to delete the phrases with vulgar or funny content as well as those having rare words. These actions will prevent future complications during the recording process, and consequently, will impact in the quality of the final recorded corpus.

Acknowledgements. This work has been partly supported by the project "Man-Machine Spoken Interaction" (LAFMI). We also want to express our gratitude to Esmeralda Uraga, Andrés González and Alberto López for their valuable help.

References

1. D. Vaufreydaz, C. Bergamini, J.F. Serignat, L. Besacier, M. Akbar, A New Methodology for Speech Corpora Definition from Internet Documents, *LREC'2000 Language Resources & Evaluation international Conference*, Athens, Greece, 2000.
2. S. Galicia-Haro. Procesamiento de Textos Electrónicos para la Construcción de un Corpus. *CORE-2003*, México, D.F. 2003.
3. Gelbukh, A., G. Sidorov and L. Chanona. Compilation of a Spanish Representative Corpus. *International Conference on Computational Linguistics and Intelligent Text Processing CICLing02*, LNCS 2276, Springer. 2002.
4. E. Uraga and L. Pineda. Automatic generation of pronunciation lexicons for Spanish. *International Conference on Computational Linguistics and Intelligent Text Processing CICLing 2002*. LNCS 2276, Springer, 2002.
5. H. E. Pérez. Frecuencia de fonemas. *Revista Electrónica de la Red Temática en Tecnologías del Habla*, Número 1, Marzo, 2003.
6. E. Alarcos-Llorach. *Fonología española*. Madrid, Gredos. 1965.