SPANISH DIALECTS: PHONETIC TRANSCRIPTION

Asunción Moreno and José B. Mariño

Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034 Barcelona, SPAIN
(asuncion/canton)@gps.tsc.upc.es

ABSTRACT

It is well known that canonical Spanish, the dialectal variant 'central' of Spain, so called Castilian, can be transcribed by rules. This paper deals with the automatic grapheme to phoneme transcription rules in several Spanish dialects from Latin America. Spanish is a language spoken by more than 300 million people, has an important geographical dispersion compared among other languages and has been historically influenced by many native languages. In this paper authors expand the Castilian transcription rules to a set of different dialectal variants of Latin America. Transcriptions are based on SAMPA symbols. The paper includes an identification of sounds that doesn't appear in Castilian, extend accepted SAMPA symbols for Spanish (Castilian) to different dialectal variants, describes the necessary rules to implement an automatic Orthographic to Phonetic transcription in several dialectal Spanish variants and show some quantitative results of dialectal differences.

1. INTRODUCTION

The study and representation of dialectal variations of a language is very important to improve the performance of recognition systems. Dialectal variations influence all the steps in any speech recognition system such as the design of training databases, the acoustic-phonetic modeling or the language modeling. An example of the importance of dialectal variations in training databases can be found in the framework of the SpeechDat project /1/ were dialectal variations of languages within and across countries are recorded i.e. different databases for German and French are recorded in those European countries were are spoken

Recently, a consortium of European Companies and public institutions have form the SALA project / 2/ (SpeechDat Across Latin America). The objective is to record Polyphone like Spanish and Portuguese Databases in Latin America to train speech recognition systems for voice driven teleservices applications.

The SALA project divides the Latin American continent in a low number of recording areas but high enough to record all the significant dialectal variants. These areas include complete countries and share dialectal phonetic similarities. Between 3000-5000 speakers will be recorded from each zone. Each zone has been divided into smaller dialectal regions, and an enough quantity of recordings must be obtained from each region.

A preliminary task for this project has been the study of phonetic dialectal differences in several Spanish dialects of Latin America to implement an automatic Spanish grapheme to phoneme transcriber . An initial version was already designed by the authors for the Castilian variant and necessary modifications have been included in the prototype. This paper includes:

- Identification of sounds that doesn't appear in the Castilian dialectal variant
- Extends the accepted SAMPA Spanish set to Latin American dialectal variants
- Orthographic to SAMPA transcription rules to the Spanish dialects spoken in the most populated dialectal areas
- Quantitative comparison of dialectal variants

2. DIALECTAL VARIANTS

A dialectal division of Latin America is a problem out of the scope of this paper. Many attempts exist to prepare dialectal Spanish atlas taking into account all the main facts that affects dialects: native languages influence, phonetics, lexical, grammar, ...

In this paper we pay attention to the phonetic point of view. It is commonly assumed that Latin America is divided in two broad phonetic categories:

- low lands: including coast lands and the Bolivia "Llanos"
- high lands: including those areas sited in the mountains that cross Latin America from Mexico to Chile.

This phonetic division can be found within most of the Latin America countries and as a result we can find stronger phonetic similarities across countries than within countries. In this paper we chose a set of six dialectal variants significant enough to represent the main dialectal phonetic variants. That are:

Mexico

We have developed the rules for the central Mexican dialect. It's the dialect spoken for the 50% of the population

Venezuela

We have developed the rules of the dialectal spoken in Caracas. It's the most populated area in Venezuela. The rules can be used in the Caribbean islands and, to some extend, in the coast of

This research was supported by the CICYT under contracts TIC95-0884-C04-02 and TIC95-1022-C05-03.

Panama and Colombia. This is a representative of “low land” dialects.

Colombia

We have developed the rules for the Spanish spoken in Bogota, the most populated area in Colombia. This is a “high land” dialect and the rules can be applied in most of the Andes zones from Venezuela to the south of Peru.

Peru

Most of the population of Peru is concentrated near the coast. We have chosen the coast phonetic variant of Lima, the capital. This dialect belongs to the “low land” category and can be applied, in a wide sense, in the South American Pacific coast from Panama to the south of Peru, although some small areas in Ecuador show very specific characteristics.

Chile

Homogeneity is one of the characteristics of the Spanish spoken in Chile. Rules are developed for the area near the capital, Santiago de Chile.

Argentina

Phonetic differences exist within this country, specially in the half northern area. We have chosen the variant of Buenos Aires. It’s applicable to the most populated area of Argentina and Uruguay

4. PHONETIC TRANSCRIPTION

The phonetic transcription of the Latin-American dialectal variants of Spanish is based on the rules for transcribing Spanish as it is spoken in the central region of Spain [3]. This initial set of rules has been modified according to the specific phonetics of every dialect [4,5].

Firstly, we took into account that the Spanish spoken in Latin-America shows two characteristics that are shared all along the continent:

a) Pronunciation of /T/ as /s/, that it is called “seseo”; and
b) /L/ is always uttered as /jj/, effect that is known as “yeísmo”.

Hence, the original set of rules was modified accordingly. Afterwards, a different transcription algorithm was obtained for each regional variant we are considering. The particular rules for each of them follow below.

Table 1 shows the sounds that are necessary to represent the Latin-American Spanish and do not exist in central Spanish. Five allophones have already been registered in the X-SAMPA inventory [6] and are represented by the standardized symbol. One sound (marked by means of *) does not form part of X-SAMPA set yet. We make a proposal to denote it following the SAMPA conventions.

Allophone	Description
dl	voiced lateral affricate *
dZ	voiced palatoalveolar affricate
h	voiceless glottal fricative
ts	voiceless alveolar affricate
C	voiceless palatal fricative
S	voiceless palatoalveolar fricative
Z	voiced palatoalveolar fricative

Table 1. Allophones added into the SAMPA Spanish set to transcribe the Latin-American Spanish. The sound marked with * is not included in the X-SAMPA inventory yet.

4.1 Mexico

The main characteristic of the Mexican dialect is the presence of sounds taken from the native language “náhuatl”. These sounds appear typically in names and words imported from this language. Table 2 gathers these sounds and provides examples.

Allophone	Example	Transcription
dl	náhua**tl**	‘na-wa**dl**
ts	que**tz**al	ke-’**ts**al
S	**x**ocoyote	**S**o-ko-’jjo-te

Table 2. “Náhautl” allophones of Mexican Spanish.

4.2 Caribbean region

The following specific rules apply:

a) The voiceless velar fricative /x/ is uttered as voiceless glottal fricative /h/.
b) When in coda position, /s/ is transformed in /h/.
c) Nasal consonants in post-nuclear position are velar /N/.

Table 3 exhibits some examples.

Allophone	Example	Transcription
h	al**g**ibe	al-‘**h**i-Be
h	pa**s**ta	‘pa**h**-ta
N	co**n**de	‘ko**N**-de

Table 3. Examples of the particular rules for Caribbean Spanish.

4.3 Colombia

The main variants of the Spanish spoken in the region close to Bogota are two:

a) The velar fricative /x/ is uttered as glottal /h/.
b) The sounds /b/, /d/ and /g/ are always pronounced as stop consonants except when coming between vowels or in post-nuclear position. In this case, the approximant allophone (/B/, /D/ or /G/) is standard.

Examples can be found in Table 4.

Allophone	Example	Transcription
B	algi**b**e	al-‘hi-**B**e
d	des**d**e	‘dez-**d**e
g	car**g**a	‘kar-**g**a

Table 4. Examples of transcription for voiced stop consonants (Spanish spoken in Colombia).

4.4 Peru

The Spanish spoken at the coast of the Pacific Ocean has the following particularities:

a) The voiceless velar fricative /x/ is produced as glottal /h/.
b) /s/ in rhyme position is realized as /h/, except at the end of a word before a pause or a vowel.
c) Nasal consonants in coda position are velar /N/.
d) In a final unstressed syllable, /D/ is omitted between vowels.

Table 5 includes some examples to illustrate these rules.

Allophone	Example	Transcription
h	re**j**as	‘rre-**h**as
s	do**s** a**s**ta**s** má**s**	‘do**s** ‘a**h**-ta**h** ‘mas
N	tie**m**po	‘tje**N**-po
D	da**d**o	‘da-o

Table 5. Specific transcription rules for the Spanish from Peru.

4.5 Chile

The following rules apply to the Spanish spoken at the central region of Chile:

a) /s/ in rhyme position is realized as /h/.
b) The voiceless velar fricative /x/ is produced as palatal /C/ when preceding the vowels /e/, /i/ or /j/.

Table 6 provides examples of transcriptions in Spanish from Chile.

Allophone	Example	Transcription
h	reja**s**	‘rre-xa**h**
x	re**j**a	‘rre-**x**a
x	co**j**ín	ko-‘**C**in

Table 6. Instances of Chilean Spanish transcription.

4.6 Argentina

The characteristic that distinguish the Spanish spoken at Buenos Aires can be summed up as:

a) The “yeísmo” becomes “zeísmo”: both /L/ and /jj/ are transformed into voiced palatoalveolar fricative /Z/. After a nasal consonant, /Z/ is produced as /dZ/.
b) /s/ in post-nuclear position is transformed in /h/, except at the end of a word before a pause or a vowel.

Table 7 shows some examples.

Allophone	Example	Transcription
Z	**ll**ave	‘**Z**a-Be
dZ	cón**y**uge	‘kon-**dZ**u-xe
s	de**s**de	‘de**h**-De

Table 7. Illustrations of the Spanish produced in Argentina.

5. QUANTITATIVE ANALYSIS

In order to quantify the relevance of the phonetic dialectal differences, a corpus has been automatically transcribed to all the dialects above mentioned and the results, in terms of relative frequency of allophones’ counts, compared.

The chosen corpus was designed in the framework of the SpeechDat project. The objective was to have speech enough to train an ASR system. The corpus fulfils the following specifications:

- The corpus is divided in sets of 9 sentences and each set contains all the allophones of the language.
- The corpus is designed to maximize the number of diphones and triphones.

The corpus contains 7200 sentences taken from spontaneous sentences, newspapers and books. It was designed to fulfil the specifications for the dialectal variant Castilian spoken in Spain. The automatic transcription of this corpus to the Castilian dialect gives an amount of 345858 allophones.

In this paper we show some of the quantitative result that we observe after a comparison of all the allophone counts in every considered dialect. We pay attention to the following effects:

- “Yeismo” and “Zeismo”: /jj/, /L/, /Z/, /dZ/
- Nasals: /n/, /m/, /N/
- Stop consonants and approximant realizations: /b/, /d/, /g/, /B/, /D/, /G/
- Production of voiceless glottal fricative and related allophones among dialects: /T/, /s/, /z/, /h/, /x/, /C/

Vowels in Spanish represent approximately a 50% of the total allophone counts and there aren’t significant differences among dialects. Every one of the last three sets has a relative frequency count of 10% over the all Spanish set. For this reason the chosen sets are significant enough.

“Yeismo” and “Zeismo” are two effects that already were commented in the above section and affects a 0.7% of the total number of allophones counts.

Figure 1 shows the differences among dialects concerning velarization of nasals. The efect of velarization is common only in the ‘coast’ dialects considered in this paper and quantitatively is very frequent because /n/ is one of the most frequent allophones in Spanish.

Figure 2 shows the quantitative effect of interchange of voiced stops, shaded dark in the figure, by approximant realization in the Coast of Venezuela, (Caribbean dialect) and the elision of approximants in the coast dialect of Peru. Very quantitative small differences can be observed in the dialectal variants of Mexico and Venezuela.

Finally Figure 3 shows the main aspects related with the voiceless glottal fricative /h/. We have grouped the following allophones: /T/, /s/, /z/, /h/, /x/, /C/. This group shows the most significant difference among “low land” and “high land” dialects and commonly is named as “loss of s”. /s/ is the most frequent consonant allophone (6.2%) in Castilian Spanish and /z/ represents 1.3% of the total allophone counts. In this figure frequency counts of /s/ and /z/ are merged together. Most of the /z/ realizations are produced as /h/ in the “low land” dialects and South Cone countries. This figure allow to observe the relative frequency counts of /x/ and it's pronunciation as /h/ in some of the dialectal variants considered.

Dialectal differences have a broader scope than the phonetic characteristics. Lexicon and grammar, also plays a very important role. A first attempt to quantify these differences is carried out in this paper. The above mentioned corpus was translated to two different dialects: Mexican and Colombian. Several native persons from both countries, Mexico and Colombia who had lived in Spain for some time and knew the Spanish spoken in Spain, were selected to modify the corpus.
Specific instructions were to modify the lexical and the grammatical structure when necessary. At the same time, the final corpus must accomplish the general SpeechDat constraint i.e. the corpus can be split in sets of nine sentences where all the allophones of the specific dialect were pronounced. The only exception to this rule was the very specific "Náhautl" allophones of Mexican Spanish because are very difficult to find in the Mexican lexicon and very often are linked with city names.

The two new generated corpus were transcribed automatically following the intended dialectal rules.

A total of 2128 sentences from the original corpus was modified to obtain the Colombian corpus.

A total of 931 sentences were modified to obtain the Mexican corpus.

6. SUMMARY

Grapheme to Phonetic transcription rules for seven different Spanish dialects have been developed and applied to an automatic grapheme to phoneme tool. The system uses an extended Spanish SAMPA set. A comparison of these dialects in terms of relative frequency counts has been done on a corpus of 350000 allophones.

7. ACKNOWLEGMENTS

The authors want to thank Elsa Mora (Universidad de los Andes, Mérida, Venezuela), Jorge Gurlekian (Laboratorio de Estudios Sensoriales, Buenos Aires, Argentina) and Miriam Cid (Pontificia Universidad Católica, Santiago, Chile) their help

8. REFERENCES

1 Hoege H. et al. (1997) European Speech Databases for Telephone Applications Proc. Int. Conf. On Acoustics, Speech and Signal Processing. ICASSP'97

2 A Moreno, H. Hoege, J. Koehler , J. B. Mariño. SpeechDat Across Latin America Project SALA. Proc. First Int. Conf. On Language Resources & Evaluation. ICLR'98

3 J. Llisterri, José B. Mariño, "Spanish adaptation of SAMPA and automatic phonetic transcription", *Report SAM-A/UPC/001/V1 (February 1993).*

4 J.M. Lipski, "*El español de América*", Cátedra, 1994.

5 M. Alvar (ed.), "*Manual de dialectología hispánica. El Español de América*", Ariel, 1996.

6 D. Gibbon, R. Moore, R. Winski (ed.), "*Handbook of Standards and Resources for Spoken Language Systems*", Mouton de Gruyter, 1997.

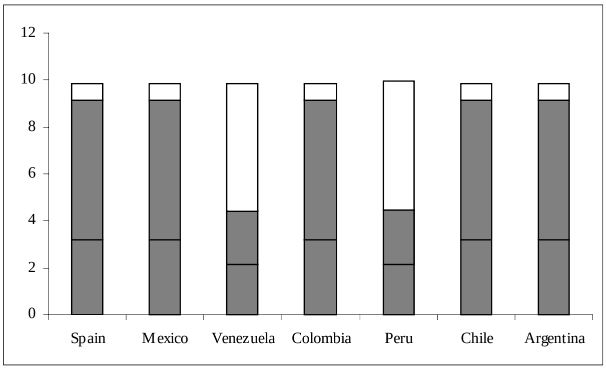

Figure 1. Comparative relative frequency counts of nasal in the considered Spanish dialectal variants. Low part: /m/, medium part: /n/, white part: /N/

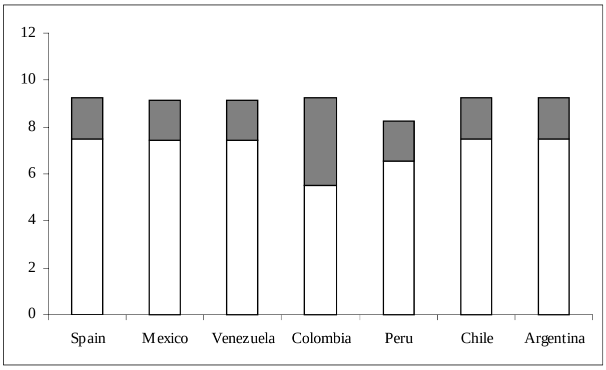

Figure 2. Comparative relative frequency counts of Stops and Approximants consonants in the considered Spanish dialectal variants: dark: /b/, /d/ /g/, white: /B/, /D/, /G/

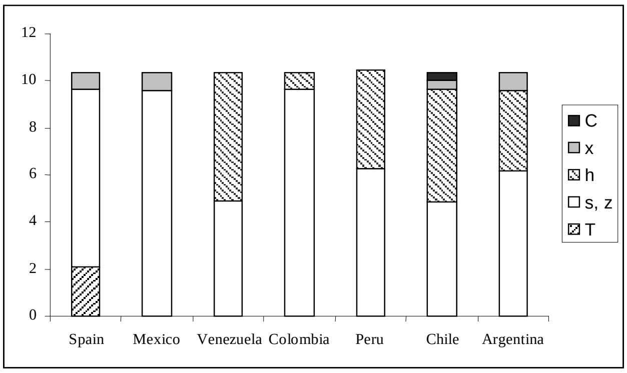

Figure 3. Comparative relative frequency counts /T/, /s/+/z/, /h/, /x/ and /C/ consonants in the considered Spanish dialectal variants.