Developing a Phonemic and Syllabic Frequency Inventory for Spontaneous Spoken Castilian Spanish and their Comparison to Text-Based Inventories

***Antonio Moren Sandoval, *Doroteo T. Toledano, *Raúl de la Torre, *Marta Garrote and **José M. Guirao**

*Universidad Autónoma de Madrid, **Universidad de Granada, Spain

antonio.msandoval@uam.es, doroteo.torre@uam.es, raul@maria.lllf.uam.es, marta@maria.lllf.uam.es, jmguirao@ugr.es

Abstract

In this paper we present our recent work to develop phonemic and syllabic inventories for Castilian Spanish based on the C-ORAL-ROM corpus, a spontaneous spoken Spanish with varying degrees of naturalness and in different communicative contexts. These inventories have been developed by means of a phonemic and syllabic automatic transcriptor whose output has been assessed by manually reviewing most of the transcriptions. The inventories include absolute frequencies of occurrence of the different phones and syllables. These frequencies have been contrasted against an inventory extracted from a comparable textual corpus, finding evidence that the available inventories, based mainly on text, do not provide an accurate description of spontaneously spoken Castilian Spanish.

1. Introduction

The first phoneme inventory with frequencies for the Spanish language was estimated by Zipf and Rogers (1939) based on the phonological description of Navarro. Since then, several studies on this topic have been produced. Table 1 summarizes previous works, with the total number of phonemes or letters used in the frequency estimation and the type of corpus (written or spoken).

Authors (year)	# phon/lett	Type
Zipf & Rogers (1939)	5,000	written
Navarro Tomás (1946)	20,000	written
Guirao & Borzone (1972)	62,980	written
Quilis & Esgueva (1980)	160,000	spoken
Rojo (1991)	3,641,915	written
Alameda & Cuetos (1995)	9,233,004	written
This study (2008)	1,244,411	spoken

Table 1: Comparison of our study and previous studies on Spanish phoneme frequencies.

For Spanish syllables, frequency studies are scarce: Álvarez, Carreiras & De Vega (1992) and Alameda & Cuetos (1995) are the most recent estimations. In the first case, 41,592 syllables were used in the computation. The latter took 3,930,954 into account. It is important to note that in both cases, the source was written texts which had not been phonologically transcribed.

In the present study, a total number of 1,244,411 phonemes and 558,982 syllables were used.

The novelty of this research is the use of a spontaneous speech corpus as source for the inventory, excluding Quilis & Esgueva (1980) who used a corpus much smaller than ours (only a tenth the size of our corpus, and only 16 speakers). For this work we have used two of the most important spontaneously spoken corpora for Spanish: CORLEC (Marin, 1992) and C-ORAL-ROM (Moreno et al., 2005). The latter is the base for the experiment described in this paper. The Spanish C-ORAL-ROM corpus consists of over 348,000 words (including prosodic marks), in 192 orthographically transcribed recordings. In total, 429 different speakers and more than 42 hours of recorded sessions. The corpus has been divided into three main classes: informal (165,210 words), formal (70,924) and media (97,170). A small subcorpus of 14,760 words is composed of telephone conversations. The quality of the transcriptions is assured by external validation (ELDA).

Section 2 presents the general methodology followed in this study. Section 3 describes the transcriptor used. Section 4 shows the main results, including a comparison of the results obtained from spoken and written corpora, and finally Section 5 exposes our conclusions on the experimental results.

2. Methodology

The C-ORAL-ROM corpus includes orthographic transcriptions for all the recordings but did not provide syllabic or phonemic transcriptions. We have generated this information in an iterative way.

In order to compile the inventory from the spoken corpus, the following steps have been taken:

1. The starting point was a transcriptor developed by our group (see next section).

2. Using the CORLEC corpus, we searched for potential transcription problems.

3. Based on the findings of the former step, new features were added to the transcriptor.

4. A first run over the C-ORAL-ROM was conducted, obtaining a preliminary phonemic and syllabic transcription.

5. The transcriptions were manually revised: 100 % were revised by one linguist and 60% of the transcriptions by two different reviewers. From these revisions the transcriptor rules (and exceptions) were improved.

6. A definitive transcription has been performed, being the base for the inventories presented in this report.

3. Transcriptor Development

The phonemic and syllabic transcriptor used for this work is based on context-dependent rewrite rules and an exception mechanism, and operates on word units. To obtain the phonemic and syllabic transcription of a word, first the word is looked up on a list of exceptions. The list of exceptions has a syllabic and phonemic transcription for each word on the list, so if the word is on the list, its associated phonemic and syllabic transcription is taken and there is nothing else to do for that word. If the word is not on the list, firstly a series of context-dependent rewrite rules are applied to obtain the phonemic transcription of the word. After that, syllables are grown from vowels by applying other set of rules. Finally, another set of rules are applied to determine whether each vowel in the word should be stressed or not according to stress and orthographic rules related to the use of the sign '´' to denote stress on Spanish vowels.

The transformation from the orthographical representation of a word to its phonemic transcription is based on context-dependent rewrite rules with the following format:

sign → [left-context(s)] new-sign(s) [right-context(s)]

where [*left-contexts(s)*] and [*right-context(s)*] are optional and may include any number of signs (which could represent letters or phonemes depending on the rule), *sign* is the letter or phoneme to rewrite and the *new-sign(s)* is zero, one or more signs representing letters or phonemes (depending on the rule). These rules are applied one-by-one in a predefined order. Given the regularity of the letter-to-sound mapping in Spanish, as few as 50 rules were enough to do a good job (as evaluated later). This was also facilitated by our decision to use a minimal phoneme set consisting of only 23 phonemes, and to consider only canonical phonemic transcriptions (not taking into account for instance regional variations or reduction phenomena). An important limitation to this transformation comes from the initial decision of considering as the input unit only a word instead of a whole sentence. This way we cannot take into account inter-word phonemic phenomena.

Once the word has been transformed into a sequence of phonemes, each vowel in the word is delimited as an initial syllable. First a couple of rewrite rules are applied to determine which pairs of vowels belong to the same syllable and which others belong to different syllables. After that, eight rewrite rules are applied to add consonants that appear before and after the vowel(s) to the syllable. If, after this process some consonants remain unassigned, that is reported as a syllabification error, which occurred mainly with foreign words and acronyms.

After the word has been transcribed phonemically and syllabified, other set of rules assigns stress to one of the syllables in the word according to the stress and orthographic conventions of Spanish and making use of the syllabification obtained in the previous phase. Here the limitation of performing the whole process on word units is again an important one because it is frequent in spoken Spanish to group several words (i.e. articles and nouns) with a single stressed syllable and our transcriptor cannot currently treat this issue.

For the development of the inventory we have tuned the rules and the exceptions of the transcriptor based on the manual corrections of the automatic transcriptions of the C-ORAL-ROM corpus. It should be noted that only 2% of the words transcribed automatically was found to have a transcription (either phonemic or syllabic) error.

For this work we have ignored the stress information, and have only taken into account the phonemic transcription and the syllabification. Thus two syllables and phonemes that differ only in stress are considered the same in this study. The exceptions included for phonemic and syllabic transcriptions correspond mainly to foreign words and acronyms for which our transcriptor produced syllabification errors.

Phn	Spanish			
	Spoken		Written	
a	152664	12.27	323783	12.89
b	31126	2,50	64170	2.55
θ	18940	1.52	50301	2.00
ʧ	3744	0.30	4463	0.18
d	54284	4.36	136187	5.42
e	188196	15.12	320140	12.74
f	6217	0.50	23042	0.92
g	11359	0.91	26138	1.04
i	89799	7.22	190756	7.59
x	7681	0.62	19362	0.77
k	55863	4.49	95427	3.80
l	56107	4.51	137148	5.46
m	39278	3.15	69445	2.76
n	87775	7.05	178012	7.09
ɲ	2427	0.19	7729	0.31
o	129208	10.38	234238	9.32
p	34135	2.74	68687	2.73
r	5236	0.42	25016	0.99
ɾ	63702	5.12	155632	6.19
s	100881	8.11	184085	7.33
t	56287	4.52	108398	4.31
u	39146	3.14	76390	3.04
ʎ	10356	0.83	13307	0.53
ALL	1244411	100	2511856	100

Table 2: Frequency of Spanish phonemes.

4. Results

In order to provide a comparison between the frequencies obtained from a written corpus and a spoken one, we selected randomly 480,000 words from a news agency corpus (EFE) with 150 million words in Spanish. The selection procedure was to choose one word every 300. This way we have a significant written corpus to compare against our spoken corpus.

The inventory extraction procedure has been the same for both corpora.

First, two word lists are extracted from each corpus: the set of forms occurring in the corpus and the same set enriched with the number of instances for every form. Both lists are subsequently fed to the transcriptor in order to obtain a phonological lexicon and a phonological corpus containing the occurrences of each word.

Spanish					
Spoken			Written		
.a.	27606	4,94	.de.	46748	4,49
.ke.	21070	3,77	.a.	37021	3,55
.de.	19638	3,51	.la.	27138	2,61
.es.	13703	2,45	.ta.	17885	1,72
.i.	13102	2,34	.ke.	17704	1,70
.no.	12781	2,28	.en.	17203	1,65
.te.	10620	1,89	.do.	16840	1,62
.el.	10282	1,84	.te.	16610	1,59
.la.	10281	1,84	.na.	15872	1,52
.do.	10172	1,82	.ma.	15463	1,48
.se.	9335	1.67	.se.	15141	1.45
.en.	8819	1.57	.to.	14614	1.40
.ta.	8726	1.56	.el.	14563	1.39
.e.	8079	1.44	.ra.	14183	1.36
.to.	7601	1.35	.ko.	13037	1.25
.si.	7535	1.34	.ka.	12657	1.21
.ko.	7395	1.32	.pa.	12482	1.19
.na.	7090	1.26	.ti.	11606	1.11
.o.	7076	1.26	.es.	10878	1.04
.ra.	6753	1.20	.kon.	9918	0.95
.lo.	6461	1.15	.por.	9795	0.94
.ba.	6329	1.13	.no.	9484	0.91
.me.	5753	1.02	.da.	9405	0.90
.ka.	5531	0.98	.los.	9403	0.90
.pa.	5492	0.98	.ba.	9159	0.88
.por.	5022	0.89	.si.	8824	0.84

Table 3: Frequency of top 25 Spanish syllables.

After this, all forms are syllabified by means of the transcriptor resulting in two sets of phonologically transcribed syllables, corresponding to the lexicon and to the corpus. The syllables in both sets are counted and ordered by frequency. This allows us to learn the syllable distribution in the lexicon and the actual syllable distribution in the corpus.

At this point, we have a table of syllables for the corpus, the actual tokens of every syllable, their frequency relative to the total and their distribution (Table 3). Due to space limitations, only the 25 most frequent syllables in the corpora are shown. The inventory is ordered by frequency in the spoken corpus.

The final step involves counting of phonemes and obtaining the total frequency for each unit. The

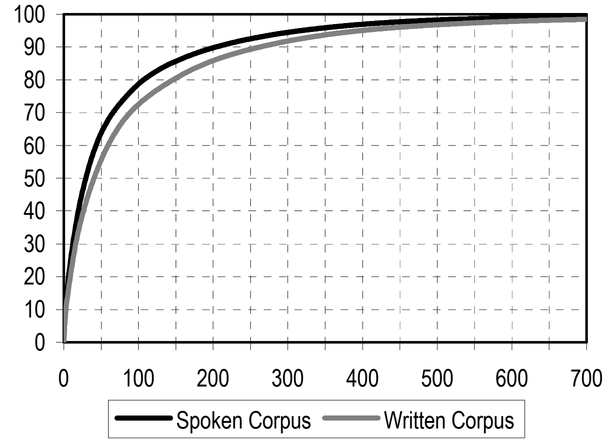

Figure 1: Distribution of syllables in the written and spoken corpus. Figure shows the cumulated relative frequency of the syllables in Spanish sorted in decreasing relative frequency.

operation is then repeated taking the syllabic context into account. Given each phoneme's frequency of occurrence, the probabilities of its occurrence in any combination can be calculated. The results for the 23 phonemes are shown in the Table 2.

Figure 1 shows the distribution of syllables both in the written and the spoken corpus. An interesting result is that the first 100 syllables represent a 78.7% of the spoken corpus while in the written corpus they represent a 72.5%. In the oral corpus the first 650 syllables cover 99.2% while in the written corpus this is 98.2%. This difference may be due to the higher lexical variety of written texts.

5. Conclusions

This is the first frequency inventory of Spanish phones and syllables based on both a spoken and a written corpus of comparable size and using the same criterion and tool for segmenting the units. Two important conclusions can be made from the experimental data:

1. Different frequency results are observed for written and spoken corpora. The order of some units and the percentage of use are different. This is especially marked in the vowels /a, e, o/. This observation suggests that training language models on written corpora could produce poorer results than training on spoken texts (provided that the amount of training spoken material is sufficient). However, it should be desirable to check that the observed differences are statistically significant and not due to differences in sampling selection.
2. With a few syllables in Spanish it can be covered a significant portion of a text. Therefore the employ of syllables instead of phonemes as basic units for acoustic modelling looks promising for Spanish.

As future work, our team will investigate further the use of syllables for automatic speech recognition system for spontaneous Spanish.

6. Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology under projects TEC2006-13170-C02-01 and TIN2007-67407-C03-02.

7. References

Alameda, J.R. & F. Cuetos (1995) Diccionario de frecuencias de las unidades lingüísticas del castellano. Servicio de publicaciones de la Universidad de Oviedo.

Álvarez, C.J., M. Carreiras & M. De Vega (1992) "Estudio estadístico de la ortografía castellana: (1) la frecuencia silábica" Cognitiva 4, pp.75-105.

Guirao, M. & A. Borzone de Manrique (1972) "Fonemas, sílabas y palabras del español de Buenos Aires" Filologia, XVI, pp.135-165.

Marcos Marín, F. (1992) "El Corpus Oral de Referencia de la Lengua Española contemporánea" Project Report. Madrid. Publisher in ftp://ftp.lllf.uam.es/pub/corpus/oral

Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, JM., & de la Torre, R. (2005) The Spanish Corpus in Cresti, & Moneglia (eds.) C-ORAL-ROM: Integrated reference Corpora for Spoken Romance Languages. Amsterdam, John Benjamins. pp. 135-161

Navarro Tomás, T. (1946) "Escala de frecuencia de los fonemas españoles" Estudios de fonología española. Syracuse, pp.15-30

Quilis, A. & M. A. Esgueva Martínez (1980)"Frecuencia de fonemas en el español hablado"Lingüística Española Actual, 2.

Rojo, G. (1991) "Frecuencia de fonemas en el español actual" en Brea, M. & F. Fernández Rei (Coords.) Homenaxe ó profesor Constantino García. Universidade de Santiago. pp.451-457.

Zipf, G. K. & J. M. Rogers (1939) "Phonemes and Variphones in four present-day romance Languages and Classical Latin from the viewpoint of dynamic Philology" Archives Néerlandaises de Phonétique Expérimentale" 15, pp. 111-147.