The Corpus DIMEx100: Transcription and Evaluation

Luis A. Pineda[a], Hayde Castellanos[a] , Javier Cuétara[b], Lucian Galescu[c], Janet Juárez[a], Joaquim Llisterri[d], Patricia Pérez[a] and Luis Villaseñor[e]

[a] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM)

[b] Facultad de Filosofía y Letras, UNAM

[c] Florida Institute for Human and Machine Cognition

[d] Departamento de Filología Española, Universidad Autónoma de Barcelona

[e] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE, Mex.)

Abstract. In this paper the transcription and evaluation of the corpus DIMEx100 for Mexican Spanish is presented. First we describe the corpus and explain the linguistic and computational motivation for its design and collection process; then, the phonetic antecedents and the alphabet adopted for the transcription task are presented; the corpus has been transcribed at three different granularity levels, which are also specified in detail. The corpus statistics for each transcription level are also presented. A set of phonetic rules describing phonetic context observed empirically in spontaneous conversation is also validated with the transcription. The corpus has been used for the construction of acoustic models and a phonetic dictionary for the construction of a speech recognition system. Initial performance results suggest that the data can be used to train good quality acoustic models.

1 Introduction

Despite recent progress in the field of speech technology, the availability of phonetic corpora for linguistic and computational studies in Spanish is still very limited (Llisterri *et al*., 2005). The creation of this kind of resources is required for a variety of reasons: TTS (Text-to-Speech) systems need to be targeted to specific linguistic communities, and acoustic models for the most common allophones of the dialect need to be considered in order to increase recognition rates in Automatic Speech Recognition (ASR) systems. Previous corpora for Mexican Spanish, like Tlatoa (Kirshning, 2001), have only considered the main phonemes of the language, and have conflicting criteria for the transcription of some consonants (e.g. *y* in *ayer*) and semi-consonant or semi-vowel sounds (e.g. [j] and [w]). Another antecedent is the SALA Corpus (Moreno et al., 2000) consisting of a set of speech files with their orthographic transcription and a pronunciation dictionary, with the canonical pronunciation of each word; this corpus is oriented to the construction of ASR for telephone applications for Mexican and other Spanish dialects. However, phonetic corpora for computational phonetic studies and spoken technology applications with a

solid phonetic foundation and a detailed phonetic analysis and transcription are much harder to find.

A linguistic and empirically motivated allophonic set is also important for the definition of pronunciation dictionaries. The phonetic inventory of Mexican Spanish, for instance, is usually described as consisting of 22 phones: 17 consonants and 5 vowels (Perissinotto, 1975), but our empirical work with the dialect of the center of the country has shown that there are 37 allophones (26 consonant sounds and 11 vowels and semi-consonants) that appear often and systematically enough in spoken language to be considered in transcriptions and phonetic dictionaries. This set needs to be further refined for the specific requirements of acoustic models in ASR (e.g. silences for unvoiced sounds). We have also observed that phonetic contexts that appear often and systematically enough can be described through phonetic rules that can be used both for theoretical studies and for the construction of speech technology applications.

In this paper we present the transcription and validation processes of the DIMEx100 Corpus (Pineda *et al.*, 2004), which was designed and collected to support the development of language technologies, especially speech recognition, and also to provide an empirical base for phonetic studies of Mexican Spanish[1]. In Section 2 we present an overview of the design and characteristics of the corpus. The sociolinguistic background of the corpus is presented in Section 3. The antecedents and definition of the phonetic alphabet, and also the variants used for the three granularities levels of transcription are described in Section 4. Section 5 deals with the phonetic distribution of the corpus, which is compared with results from previous studies. In Section 6 we discuss the extent to which the DIMEx100 Corpus satisfies a set of phonetic rules defined empirically in a previous study for Mexican Spanish (Cuétara, 2004). Section 7 is devoted to assessing the potential for the DIMEx100 Corpus to be used for training acoustic models for speech recognition. We conclude with a discussion about the contribution of the present work.

2 Corpus design and characteristics

For the collection process the Web was considered as a large enough, complete and balanced, linguistic resource, and the corpus sentences were selected from this source; the result of this exercise was Corpus230 (Villaseñor *et al.*, 2004), a collection of 344K sentences with 236K lexical units, and about 15 million words. From this original resource we selected 15,000 sentences with length ranging from 5 to 15 words; these sentences were ordered according to their perplexity[2] value from lowest to highest, and we retained the 7000 sentences with the lowest value. Sentences with foreign words and unusual abbreviations were edited out, and the set was also edited for facilitating the reading process and for enhancing the relationship between text and sound (e.g. acronyms and numbers were spelled out in full). The final result was a

[1] http://leibniz.iimas.unam.mx/~luis/DIME/

[2] Perplexity is a commonly used measure of the goodness of a language model that could be intuitively thought of representing the average number of word choices at every predictive step; the lower the number, the better.

set of 5010 sentences. For recording the corpus, we recruited 100 speakers; each recorded 50 individual sentences. The remaining 10 sentences were recorded by all 100 speakers; this data was collected in order to support experiments involving a large set of speakers given the same phonetic data, like speaker identification and classification. Thus, the spoken data collected included a total of 6000 sentences: 5000 different sentences recorded one time and 10 sentences recorded 100 times each. The final resource has been named the DIMEx100 corpus. In order to measure the appropriateness of the corpus we controlled the characteristics of the speakers, as described in Section 3; we also measured the frequency of occurrence and the distribution of samples for each phonetic unit, and verified that these were complete in relation to our allophonic set and balanced in relation to the language. These figures are presented below in this paper.

The corpus was recorded in a sound studio at CCADET, UNAM, with a Single Diaphragm Studio Condenser Microphone Behringe B-1 and a Sound Blaster Audigy Platinum ex (24 bit/96khz/100db SNR) using the WaveLab 4.0 program[3]; the sampling format is mono at 16 bits, and the sampling rate is 44.1 khz.

The transcription process was carried on by expert phoneticians. A basic phonetic alphabet including 54 units was used (T-54). This process was supported by an automatic transcriber that provided canonical pronunciations of each word in terms of a set of grapheme to phone rules, and also default durations for each unit (Cuétara, 2004; Pineda *et al*., 2004). The default transcription was inspected by phoneticians who carefully reviewed the pronunciation of each word, and provided the transcription of its actual phonetic realization. The transcription was time-aligned, and careful attention was paid to the determination of the boundaries of each allophonic unit. In addition to this fine transcription, two additional transcriptions were produced: T-44 and T-22, with 44 and 22 units respectively, as will be explained below. In order to facilitate building a phonetic dictionary with allophonic variation for each granularity level, the orthographic transcription of each word was time-aligned with its phonetic realization, so that all realizations of the same word in the corpus could be collected automatically.

3. Sociolinguistic considerations

Recording a spoken corpus implies considering and designing minimal linguistic measurable aspects in order to be able to evaluate them afterwards. Following Perissinotto's (1975) guidelines, speakers were selected according to age (16 to 36 years old), educational level (with studies higher than secondary school) and place of origin (Mexico City). A random group of speakers at UNAM (researchers, students and teachers) brought in a high percent of these kind of speakers: the average age was 23.82 years old; most of the speakers were undergraduate (87%) and the rest graduate, and most of the speakers (82%) were born and lived in Mexico City. As we accepted everyone interested (considering that Mexico City's population is representative of the whole country), 18 people from other places residing in Mexico City participated in the recordings. The groups of speakers was gender balanced (49% men and 51%

[3] http://www.steinberg.net/

women). Although Mexican Spanish has several dialects (from the northern region, the Gulf Coast and Yucatan's Peninsula, to name only a few) Mexico City's dialect represents the variety spoken by most of the population in the country (Canfield, 1981; Lope Blanch, 1963-1964; Perissinotto, 1975).

4. Phonetic alphabet and granularity of transcription

From a computational perspective, Mexican Spanish has been the subject of very few number of phonetic studies; in this context, the transcription of a large, high-quality corpus faced two problems: the definition of an appropriate computational phonetic alphabet and the identification of the allophonic set useful for computational applications. There are antecedents of phonetic alphabets for this dialect of Spanish from both the European and American traditions – i.e. SAMPA (Wells, 1998) and Worldbet (Hieronymus, 1997) respectively. SAMPA was originally defined for Castillian Spanish, and although it was extended to six American dialects within the context of the SALA project[4], the effort was centered in formalizing the sounds with indigenous roots (Moreno and Mariño, 1998). Later on the same authors proposed an inventory of phones and allophones of American Spanish (Moreno et al., 2000). Worldbet, on its part, does include a version for Mexican Spanish (Hieronymus, 1997) but this is exactly the same as the one listed for Castillian Spanish; consequently, this version considers two phonemes that are only part of Castilian Spanish (the fricative [T] and the lateral palatal [L]) but, on the other hand, it leaves out many allophones that are common in Mexican Spanish, like the palatalized unvoiced stop [k^j], the unvoiced dentalized fricative [s̪], the alveolar voiced fricative [s̬], the approximants, some vowel sounds, like palatalized central open [a^+], the velarized central open [a̠], the mid velar opened [o̞], among others. Another alphabet within the American tradition is the Oregon Graduate Institute Alphabet OGIbet (Lander, 1997) which also has a Mexican version (Kirshning, 2001); however, this only considers the main phonemes of the language, and has conflicting criteria for the transcription of some consonants; for instance, the palatal [Z] is considered in OGIbet as a glide, when it is in fact a consonant sound. Also, this alphabet confuses the paravocal forms of the vowels [i] and [u] with consonant sounds, and is not specific enough for the taps and trills (three different sounds are proposed but there should be only two). For a very comprehensive discussion of computational phonetic alphabets for Mexican Spanish see Cuétara (2004).

We started the DIME Project (Villaseñor *et al*., 2000, Pineda *et al*. 2002) with the goal of identifying empirically a set of allophones for Mexican Spanish that would also be appropriate for the development of spoken language technologies. As a result, the Mexbet alphabet was proposed (Cuétara, 2004). This phonetic alphabet specifies a set of 37 allophones (26 consonant and 11 vowel sounds, as shown in Tables 1 and 2 respectively), occurring often and systematically enough, and can be clearly

[4] SALA includes a speech corpus of Mexican Spanish with orthographic transcriptions and a pronunciation lexicon with a phonemic transcription (i.e. canonical pronunciations), and it is targeted for the construction of ASR Systems for mobile telephone applications. SALA is available as an ELRA resource at: http://catalog.elra.info/index.php.

distinguished using acoustic and phonetic criteria. For practical reasons, the notation of Mexbet is based on Worldbet. Mexbet was used as the main reference for the transcription of the DIMEx100 Corpus. The equivalence between Mexbet and IPA is shown in Appendix 5.

Consonants	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar
Unvoiced stops	[p] *papá*		[t] *Tío*		[k_ j] *queso, kilo*	[k] *cama*
Voiced stops	[b] *van, bien*		[d] *diente, un día*			[g] *gato, un gato*
Unvoiced affricate					[tS] *hacha*	
Voiced affricate					[dZ] *lluvia, yunque un yunque*	
Unvoiced fricatives		[f] *foco*	[s_ [ ] *Asta*	[s] *sol, cielo*		[x] *paja, geranio*
Voiced fricatives				[z] *mismo*	[Z] *ayer, el yunque*	
Approximants	[V] *haba*		[D] *Hada*			[G] *el gato*
Nasals	[m] *más*		[n_ [ ] *Antes*	[n] *nene*	[n~] *año*	[N] *angel*
Lateral				[l] *loco*		
Tap				[r( ] *pero*		
Trill				[r] *perro*		

Table 1. Consonant sounds

Vowels	Palatal				Cent.	Velar				
Semi-vowels / semi-consonants	[j] *viene, hay*									[w] *suave, aura*
Close		[i] *ahí*							[u] *su*	
Mid			[e] *meta*					[o] *lo*		
				[E] *erre*			[O] *sol*			
Open				[a_ j] *aire*	[a] *la*	[a_2] *aunque, alma*				

Table 2. Vowel sounds

In addition to the basic set, Mexbet includes a number of symbols useful for language technologies, in particular, for codifying the silences of unvoiced sounds, for

marking stressed vowels and also non-contrasting sounds in syllabic coda, which correspond to archiphonemes in traditional phonological studies.

In this study we also intended to explore the impact of transcription granularity. The granularity of a phonetic alphabet constrains the wealth of phonetic phenomena that can be studied with such an alphabet. In particular, an alphabet with 22 symbols (phonemes) permits to express very few pronunciations for words and limits strongly the variety of phonetic contexts that can be studied. However, the availability of the Mexbet alphabet and the wealth of phonetic information of the DIMEx100 Corpus, permitted us to study allophonic variation systematically. To this end, we transcribed the corpus at three levels of granularity, which we called T-54, T-44 and T-22 according to the number of phonetic units included for each level (i.e., 54, 44 and 22 units, respectively).

The T-54 level is used for narrow transcriptions, and includes the allophonic set in Table 1, in addition to the closures of the 8 unvoiced sounds and 9 stressed vowels, as shown in Appendix 1. Spanish is a free stress language; for instance, the words *número* (number) *numero* (I enumerate) and *numeró* (he/she enumerated something) have very different meanings. Since there are acoustical and perceptual differences between stressed and unstressed vowels (Llisterri *et al.*, 2003) we are interested in assessing the effects on recognition performance due to variations in duration; another parameter that affects significantly the length of a vowel is whether the segment is open or closed. Although a detailed analysis these data is still pending, Appendix 4 shows the durations in milliseconds, together with the standard deviation, for all allophones at all three levels of transcription.

The T-44 level is a broader transcription, including the basic allophonic set (17 consonants and 5 vowels), 7 closures of stop consonants, 3 approximant sounds ([V, D, G]), 2 semi vowels or semi consonants ([j] and [w]) and 5 stressed vowels; in addition, this level includes 5 special symbols to subsume consonants sounds in syllabic codas, that have no contrasting values in Spanish (Quilis, 1981/1988); these are /p – b/, /t – d/, /k – g/, /m – n/ and /r(– r/ and are represented by [-B], [-D], [-G], [-N] and [-R] respectively. The full T-44 set is shown in Appendix 2.

The T-22 level corresponds to the basic set of 17 consonants and 5 vowels of Mexican Spanish, as shown in Appendix 3. As was mentioned, the transcription process of the T-54 level has been supported by a tool that produced a basic time-aligned transcription of the standard pronunciation of the words, by means of a set of grapheme to phone transcription rules (Cuétara, 2004; Pineda *et al.*, 2004). However, the final representation of each unit, as well as the specification of its time boundaries, was the result of decisions made by expert phoneticians. The T-44 and T-22 levels were produced automatically from the T-54 through suitable Perl scripts, although the syllabic codas of the T-44 level were also manually tagged.

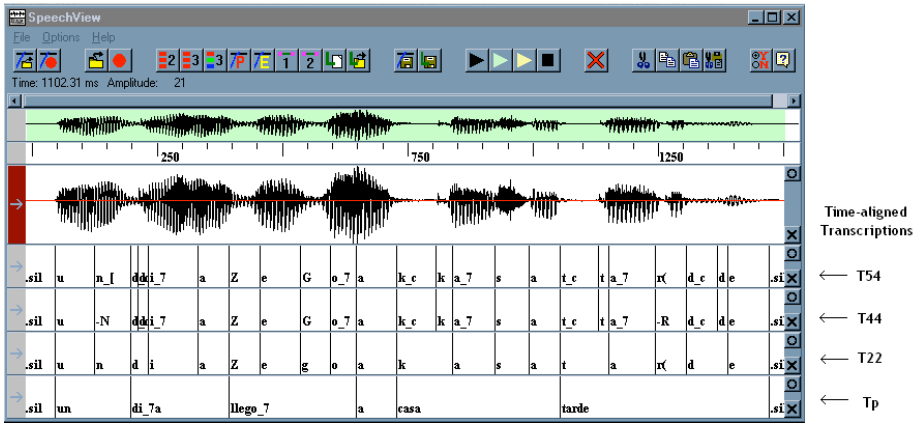

Figure 1. Time-aligned transcriptions of T-54, T-44, T-22 and word levels

In addition to these three phonetic levels, a fourth, lexical level with the time-aligned orthographic transcription of all words was produced manually. Words follow the standard Spanish orthography, with the exception of diacritics for stressed vowels, which are specified as a postfixed "_7", and the diacritic for *ñ* which is specified as "n~", reflecting the corresponding phonetic transcription. This convention was designed to allow processing with ASCII-only tools; the orthography can be easily transformed into other encodings. An illustration of the transcription of a corpus sentence with all four time-aligned transcriptions is shown in Figure 1. For the transcription process the Speech View tool was used (Sutton *et. al.*, 1998).

The time-aligned transcription of the three granularity levels with the orthographic transcription permitted the automatic collection of a phonetic dictionary for each level, including all realizations of each word in the corresponding level. As expected, a word may have several pronunciations, and the narrower the transcription level the higher the number of pronunciations for a given word. Some examples of transcription at the three granularity levels are shown in Table 3.

Word	T-22	T-44	T-54
acción ***action***	a g s i o n a k s i n	a -G s j o_7 -N a_7 -G s i -N	a G s j o_7 n a_7 k_c k s i n
conferencia ***conference***	k o f e r(e n s i a k o f e r(e n s i e k o n f e r(e n s i a	k_c k o f e r(e_7 n s j a k_c k o f e r(e_7 n s j e k_c k o n f e r(e_7 n s j a	K_c k o f e r(e_7 n s j a k_c k O f e r(e_7 n s j e k_c k o n f e r(e_7 n s j a k_c k O n f E r(e_7 n s j a
hasta ***until***	a s t a s t a a s t e s t a	a_7 s t_c t a_7 s t_c t a a_7 s t_c t e a s t_c t a s t_c t a a s t_c t e s t_c t a	A_7 s_[t_c t a_7 s_[t_c t a a_7 s_[t_c t E a s_[t_c t a s_[t_c t a a s_[t_c t a_2 a s_[t_c t E s_[t_c t a
desarrollo ***development***	d e s a r o i d e s a r o Z d e s a r o Z o d e s a r(o Z o d e s a r o Z u	D e s a r o_7 Z d_c d e s a r o_7 Z o D e s a r o_7 Z o D e s a r(o_7 Z o d_c d e s a r o_7 Z w	D e s a r O_7 Z D e s a r o_7 dZ_c dZ o D e s a r o_7 Z o D e s a r(o_7 Z o D e s a r O_7 Z o d_c d e s a r o_7 dZ_c dZ o d_c d e s a r O_7 Z o d_c d e s a r O_7 Z w
ciencias ***sciences***	s i e n s i a s i e n s i a s s i e s i a s s i e s i e s s i n s i a s s i n s i e s	s j e_7 -N s j a s i_7 -N s j e s s i -N s j a s s j e_7 -N s j a s s j e_7 s j a s s j e_7 s j e s s j e -N s j a s	s i_7 n s j e s s i n s j a s s j e_7 n s j a s j e_7 n s j a s s j e_7 n s j a z s j e_7 s j a z s j e_7 s j E z s j e n s j a s s j e n s j a z

Table 3: Different word pronunciations in levels T-22, T-44 and T-54

5. Phonetic distribution

When the corpus was originally collected, the text-to-phone translation rules allowed us to evaluate whether the corpus was complete, large enough and balanced. As an initial exercise, we translated the text into its phonemic and allophonic representations, and computed the number and distribution of samples, as reported in Pineda *et al.* (2004). However, with the transcription of the full corpus, we have been able to compute the actual statistics, as shown in Table 4. As expected, the corpus includes all phonetics units at the three granularity levels, with a large number of instances for each unit. In particular, the less represented phonetic units are [n~] with 346 samples, [g] with 426 and [dZ] with 126. Since we have a significant number of instances of all allophones in the corpus, we conclude that the corpus is complete. This is consistent with the perplexity-based method used for the corpus design, despite that this computation was performed at the level of the words.

Unit	Instances	Percentage	Unit	Instances	Percentage
p	6730	2.62%	l	14058	5.48%
t	12246	4.77%	r(	14784	5.76%
k	8464	3.30%	r	1625	0.63%
k_j	1285	0.50%	i	9705	3.78%
b	1303	0.51%	i_7	3941	1.54%
V	4186	1.63%	j	8349	3.25%
d	3881	1.51%	e	23434	9.13%
D	10115	3.94%	e_7	6883	2.68%
g	426	0.17%	E	3083	1.20%
G	1899	0.74%	E_7	1153	0.45%
tS	385	0.15%	a	18927	7.38%
f	2116	0.82%	a_7	8022	3.13%
s	20926	8.15%	a_j	539	0.21%
s_[	2912	1.13%	a_j_7	228	0.09%
z	2123	0.83%	a_2	1277	0.50%
x	1994	0.78%	a_2_7	1164	0.45%
Z	720	0.28%	o	15088	5.88%
dZ	126	0.05%	o_7	4200	1.64%
m	7718	3.01%	O	3064	1.19%
n	12021	4.68%	O_7	1533	0.60%
n_[	4899	1.91%	u	3431	1.34%
N	848	0.33%	u_7	1716	0.67%
n~	346	0.13%	w	2752	1.07%

Table 4. Phonetic distribution of the T-54 level (without closures)

These figures can also be used to assess whether the corpus is balanced. In Table 5 we compare the distribution of the DIMEx100 corpus in the T-54 transcription level to the distribution reported by Llisterri and Mariño (1993) for Peninsular Spanish. As can be seen, our balancing procedure produced figures that resemble the figures of previous studies very closely, taken into account allophonic differences between the dialects. In particular, the correlation at the level of the phones between DIMEx100 and Llisterri and Mariño (1993) is 0.98; for all this, we conclude that DIMEx100 is fairly balanced. Further data on the frequency of occurrence can be found in Navarro Tomás (1946), Alarcos (1950), Quilis and Esgueva (1980) and Rojo (1991) for Peninsular Spanish, in Guirao and Borzone (1972) for Argentinian Spanish and in Pérez (2003) for Chilean Spanish.

Phones	Phonetic Units	L & M (1993)	T-54
/p/	[p]	2.60	2.62
/t/	[t]	4.63	4.77
/k/	[k]	4.04	3.30
	[k_j]	-	0.50
/b/	[b]	0.45	0.51
	[V]	2.47	1.63
/d/	[d]	0.76	1.51
	[D]	3.20	3.94
/g/	[g]	0.11	0.17
	[G]	0.79	0.74
/tS/	[tS]	0.40	0.15
/f/	[f]	0.51	0.82
/T/[6]	[T]	1.53	-
/s/	[s]	6.95	8.15
	[s_[]	-	1.13
	[z]	1.33	0.83
/x/	[x]	0.63	0.78
/Z/	[Z]	0.19	0.28
	[dZ]	-	0.05
/m/	[m]	3.63	3.01
/n/	[n]	7.02	4.68
	[n_[]	-	1.91
	[N]	0.46	0.33
/n~/	[n~]	0.27	0.13

Phones	Phonetic Units	L & M (1993)	T-54
/l/	[l]	4.25	5.48
/L/[5]	[L]	0.54	-
/r(/	[r(]	4.25	5.76
/r/	[r]	0.40	0.63
/i/	[i]	4.29	3.78
	[i_7]	-	1.54
	[j]	2.60	3.25
/e/	[e]	13.73	9.13
	[e_7]	-	2.68
	[E]	-	1.20
	[E_7]	-	0.45
/a/	[a]	13.43	7.38
	[a_7]	-	3.13
	[a_j]	-	0.21
	[a_j_7]	-	0.09
	[a_2]	-	0.50
	[a_2_7]	-	0.45
/o/	[o]	10.37	5.88
	[o_7]	-	1.64
	[O]	-	1.19
	[O_7]	-	0.60
/u/	[u]	1.98	1.34
	[u_7]	-	0.67
	[w]	1.35	1.07

Table 5. Phonetic distribution

6. Phonetic analysis

Spanish phonetic allophonic contexts that are frequent and systematic enough can be modeled through phonetic rules. This information can be useful for phonetic studies and has potential applications in language technology; for instance, for the creation of pronunciation dictionaries for ASR, for the definition of grapheme-to-phone conversion rules with allophonic variation, or for producing more natural speech synthesis. As was mentioned, from an empirical study of the DIME Corpus, and following general studies of the phonetics of Mexican Spanish (e.g. Moreno de Alba, 1994), Cuétara (2004) verified common allophonic forms of each phone. Although most of these data are well known for the language, in the present study we report the actual figures in the DIMEx100. The counts of these contexts with their frequencies are shown in Table 6. This table presents the phoneme and a number of relevant reference contexts in which specific allophonic variation can occur. Contexts are represented by "_{…}" or "{…}_" where "_" indicates the position of a specific

[5] /L/ is part of Castilian Spanish phonological inventory only.
[6] /T/ is part of Castilian Spanish phonological inventory only.

allophonic form, the filler, and the ellipsis represents a disjunction of possible allophones, the reference context. The symbols "///_" and "_$" signal absolute start and ending respectively. The third column shows the total number of instances of the reference context that appear in the whole of the DIMEx100 Corpus. The possible fillers with their corresponding frequency (up to three) are shown in the right columns of the table. For instance, Cuétara confirmed that an allophonic palatalized form of the phone /k/, represented [k_j], precedes very often the vowels /i/ and /e/ and the semivowel /j/, but the velar form occurs elsewhere; as can be seen in Table 6, the allophone [k_j] (with its closure) do precedes the context "_{e, i, j}" 83% of the times, and the velar form [k] occurs the remaining 17% of the times in this context; on the other hand, the palatal form occurs 5% of the times in any other context, where the velar stop occurs the rest of the times (95%). As a second example consider the contexts for the bilabial voiced stop /b/; although the initial /b/ (absolute or after a pause) occurs very seldom after a pause (159 total instances) 96.86% of the times is a stop, but the approximant [V] also occurs in these initial contexts (3.14%). This distribution pattern for the stop and approximant forms of /b/ also occurs following [m] or [n], although the pattern "{m,n}_" occurs much more often (1,438 instances). The table also shows that in other contexts, out of 14,628 instances, the stop occurs 14.84% and the approximant 85.16%. It is interesting to note that the ratio of stops and approximants in similar contexts also holds for the dental and velar voiced stops /d/ and /g/, and also for the palatal voiced fricative /Z/, where the closure in these three contexts is lost most of the times, except in starting position or after [m] or [n]. As a final illustration consider the contexts of interest for the alveolar fricative /s/ phone. As noticed by Navarro Tomás since his seminal work (1918:107), the voicing of /s/ occurs only 1.54% of the times. However, /s/ is realized as a voiced sound when it precedes a voiced stop, the voiced palatal fricative, a nasal, a tap or a trill (66.26%) but it remains unvoiced the remaining times in these contexts. Also, the dental sound (i.e. s_[) appears almost always preceding a dental stop. Finally, in other contexts, the unvoiced fricative appears 89.58% of the times, the voiced form 4.64% and the dental form 5.77%. The contexts for the remaining phonemes are also shown in Table 6. Phonemes not listed have only one allophonic form, which occurs most of the time.

Phone	**Reference Context**	**Units**	**Allophones and frequencies**					
Velar unvoiced stops k			**Filler**	**%**	**Filler**	**%**	**Filler**	**%**
/k/	_{e, i, j}	3,032	k_c k_j:	83.00	k_c k:	17		
/k/	Elsewhere	25,430	k_c k_j:	5.00	k_c k:	95		
Bilabial voiced stops b								
/b/	///_	159	B_c b:	96.86	V:	3.14		
/b/	{m, n}_	1,438	B_c b:	97.91	V:	2.09		
/b/	Elsewhere	14,628	B_c b:	14.84	V:	85.16		
Dental voiced stops d								
/d/	///_	549	D_c d:	98.72	D:	1.28		
/d/	{m, n}_	3,498	D_c d:	99.26	D:	0.74		
/d/	Elsewhere	36,132	D_c d:	19.25	D:	80.75		

Velar voiced stops g								
/g/	///_	48	G_c g:	97.92	G:	2.08		
/g/	{m, n}_	384	G_c g:	76.56	G:	23.44		
/g/	Elsewhere	6,488	G_c g:	13.70	G:	86.30		
Palatal voiced fricative Z								
/Z/	///_	40	dZ_c dZ:	90.00	Z:	10		
/Z/	{m, n}_	34	dZ_c dZ:	47.06	Z:	52.94		
/Z/	Elsewhere	2,764	dZ_c dZ:	10.49	Z:	89.51		
Alveolar unvoiced fricative s								
/s/	V_V	10,988	z:	1.54	s_[:	0.06	s:	98.40
/s/	_{b, d, g, Z, m, n, l, r, r(}	5,732	z:	66.26	s_[:	0.66	s:	33.08
/s/	_{t}	5,754	z:	0.00	s_[:	99.76	s:	0.24
/s/	Elsewhere	51,083	z:	4.64	s_[:	5.77	s:	89.58
Nasal alveolar n								
/n/	_{t, d}	9,762	n_[:	99.88	N:	0.02	n:	0.10
/n/	_{k, g}	1,642	n_[:	0.49	N:	96.10	n:	3.41
/n/	Elsewhere	41,482	n_[:	11.88	N:	2.33	n:	85.79
Palatal close vowel i								
/i/	_{a, e, o, u}	7,982	j:	90.60	i:	9.40		
/i/	{a, e, o, u}_	1,451	j:	82.80	i:	17.20		
/i/	Elsewhere	21,888	j:	38.00	i:	62.00		
Palatal mid vowel e								
/e/	_{r}	381	E:	56.40	e:	43.60		
/e/	{r}_	1,149	E:	63.30	e:	36.70		
/e/	_{p, t, k, b, g, d, tS, f, x, Z}$	95	E:	27.40	e:	72.60		
/e/	Elsewhere	33,898	E:	12.40	e:	87.60		
Open vowel a								
/a/	_{u, x}	1,039	a_2:	73.60	a:	24.80	a_j:	1.50
/a/	_{l}$	357	a_2:	98.00	a:	2.00	a_j:	0.00
/a/	_{tS, n~, Z, j}	623	a_2:	1.30	a:	14.60	a_j:	84.10
/a/	Elsewhere	24,105	a_2:	9.90	a:	87.20	a_j:	2.90
Velar mid vowel o								
/o/	_{r}	209	O:	47.80	o:	52.20		
/o/	{r}_	174	O:	49.40	o:	50.60		
/o/	_{consonant}$	1,346	O:	37.30	o:	62.70		
/o/	Elsewhere	22,235	O:	20.50	o:	79.50		
Velar close vowel u								
/u/	_{a, e, o, i}	1,918	w:	97.00	u:	3.00		
/u/	{a, e, o, i}_	1,055	w:	84.70	u:	15.30		
/u/	Elsewhere	7,879	w:	34.90	u:	65.10		

Table 6. Phonetic contexts and allophonic frequencies

7. Phonetic information for speech recognition

In order to test the quality of the phonetic data for use in speech recognition applications, we built acoustic models at the three transcription granularity levels and assessed recognition performance. The data for these experiments consisted of the 5,000 utterances in the DIMEx100 Corpus recorded by 100 different speakers (the 10 common utterances that were recorded by all 100 speakers were not used). To allow meaningful comparisons, the same data was used for training and testing the acoustic models and the language models at the three transcription levels.

We assessed recognition performance for unseen data by cross-validation, using part of the corpus for training acoustic and language models and the remaining data for testing. We partitioned the data by speakers, such that no test data from a particular speaker was used for training the acoustic models.

For performing speech recognition experiments, we used the Sphinx speech recognizer (Sphinx, 2006). For alignment and scoring we used NIST's SCLITE version 1.5 package (NIST, 2005).

Acoustic Models

Well-trained broad-coverage acoustic models (AMs) typically require hundreds of hours of audio data; such volume of data makes it possible to use un-aligned transcriptions. This form of unsupervised training is clearly suboptimal, since it is practically impossible to know for a particular word instance precisely what pronunciation is used; in fact, pronunciation dictionaries used for automatic alignment commonly include just the most common pronunciation for each word. Nonetheless, the technique is quite attractive because the performance-to-cost ratio is excellent. The DIMEx100 Corpus is not large enough to be used by itself for acoustic modeling for, say, the broadcast news transcription domain, but it could be used as an additional resource; plus, it offers the opportunity to study the use of fine-grained phonetic distinctions in the phoneset. Based on the counts for phonetic unit instances shown in Appendix 4, we judged that the corpus is sufficiently comprehensive, and therefore suitable for training reasonably good acoustic models. We used the freely available SphinxTrain software package version 3.4 (Sphinx, 2006) to train context-dependent triphone models based on a 3-state continuous Hidden Markov Model architecture with 8 Gaussians per state. The complete phone set included two additional special phones, one for recognizing silence and one for background noise; these models are used by the speech recognizer to discriminate speech from non-speech in the acoustic signal. Although great attention was placed in the annotation of phonetic boundaries in the manual transcription, this information was not used in the present experiments; instead, we relied on SphinxTrain's automatic time alignments. We leave it as a future exercise to verify the agreement between the automatic time alignments and the manual ones, as well as to compare the recognition performance achieved with AMs trained on manual alignments vs. automatic alignments.

We counted the numbers of diphone and triphone types and instances in the DIMEx100 Corpus for each level of transcription, and also identified the diphones and triphones that have a large frequency. These counts are shown in Table 7 for four data points. The number of types for both diphones and triphones increases very

slowly with the amount of data. Also, the number of types of high frequency diphones and triphones (the two frequency thresholds considered were 0.5% and 0.1%) appears to have stabilized after seeing only 25% of the data. These figures suggest that further increases in the amount of data would yield only a small number of new types with significant frequencies, and the AMs would be enriched only marginally with a larger amount of corpus data.

Corpus portion	Diphones				Triphones			
	Instances	Types	>0.5%	>0.1%	Instances	Types	>0.5%	>0.1%
	T-22 Transcription Level							
25%	61K	361	68	181	59K	2904	12	239
50%	126K	385	67	183	122K	3343	12	231
75%	193K	397	69	184	187K	3567	10	232
100%	263K	413	69	185	254K	3778	12	234
	T-44 Transcription Level							
25%	61K	839	49	244	59K	6404	6	147
50%	126K	913	47	248	122K	8046	4	139
75%	195K	967	49	246	189K	9075	5	142
100%	265K	1027	49	242	256K	9835	5	144
	T-54 Transcription Level							
25%	61K	1198	36	252	59K	8884	3	107
50%	126K	1343	38	249	122K	11589	3	110
75%	195K	1413	39	249	189K	13359	3	116
100%	265K	1481	40	246	257K	14716	3	114

Table 7. Diphones and triphones statistics in the DIMEx100 Corpus

Lexicon

The full corpus includes 8,881 word types with a total of 51,893 word tokens or occurrences. Some words have multiple pronunciations in the corpus. Due to increased specificity in the transcription of allophones, the number of word pronunciations varies dramatically with transcription granularity. Thus, whereas for level T-22, we have on average 1.28 pronunciations per word, this number increases to 1.64 at level T-44 and to 1.97 at level T-54. The reason for this is that while a coarse phonetic alphabet subsumes diverse pronunciation phenomena in the units available, a finer transcription permits to account for a large set of pronunciation subtleties.

It might be tempting to include all these pronunciation variations to the speech recognition models; however, if done indiscriminately, this will also have the effect of increasing confusability and therefore generating more recognition errors[7]. For a discussion of methods to model pronunciation variation in speech recognition systems, see (Strik and Cucchiarini, 1998). For the experiments reported here, we decided to use only one pronunciation per word (the most frequent one in the training

[7] Indeed, we verified experimentally that word recognition performance on unseen data may be up to 50% worse when all pronunciation alternatives are included in the dictionary.

data for each model); we leave it as further work to study in more detail how to use alternative pronunciations to improve speech recognition performance.

Language Models

We trained trigram language models (LMs) with Witten-Bell discounting using the CMU-Cambridge Statistical Language Model Toolkit version 2.05 (Sphinx, 2006). One problem we have to take into account is the presence of out-of-vocabulary (OOV) words, that is, words present in the test data that were not seen in the training data. The literature suggests that each OOV word may produce up to two to three word recognition errors (Fetter, 1998). To insure that LMs have good lexical coverage, as well as good n-gram coverage, a good option is to collect as much textual data as possible to use in training. Our goal here is, however, not to produce a good, generic speech recognition system, but simply to validate that the DIMEx100 Corpus is useful for training acoustic models for such systems; for this reason we constructed minimal LMs with the data available in the Corpus instead of using richer LMs, as the resulting increase in SR performance due to better language modeling might obscure the contribution of the acoustic models.

Experimental Results

We performed 100-fold cross-validation, using data from a single speaker as test data for each fold. However, due to the onerous time and resource requirements for such a large experiment, we decided to use the same AMs for every 10 folds; thus, for every fold, only 90% of the data is used to train the AMs, and 99% of the data is used to train the LMs. Even so, the OOV rate remains very high, at an average of 10.1%, which is sure to have a very significant impact on recognition performance. Indeed, as shown in Table 8, average word error rates are above 30% for all transcription levels. A more detailed analysis of the errors reveals, however, that close to two thirds of them appear in proximity to OOV words. If we look at WER rates only on segments without OOV words – these segments were identified based on alignments between hypothesis and reference utterances by eliminating contiguous regions of errors corresponding to at least one OOV word – we see much better results, as shown in the WER(I) column. In fact, these results are quite good, considering the low quality of the LMs. Indeed, the average LM perplexity is 316; as a comparison, perplexity values for very large vocabulary trigram models for English, for which the literature is more abundant, are typically just above 100. We should also note here that the segments with OOV words cover just 16-17% of the data, which indicates that the effect of each OOV word on the WER was much lower than we had expected (at most 1.5 word errors per OOV word, on average). Conversely, this also means that the WER(I) estimates are not overly optimistic.

Trans. Level	WER	WER(I)
T-22	32.27	12.6
T-44	33.65	15.0
T-54	34.04	15.1

Table 8. Speech recognition performance results

Although we do see a slight decrease in performance for the finer transcription levels, we are encouraged by the fact that it is rather small, since the inclusion of more allophones is bound to increase phone confusability. It remains to be seen if further tuning of the acoustic model training process will yield even better results.

Finally, the larger number of phonetic units in the finer-grained AMs doesn't incur a significant computational cost. The average recognition time increased by just 5% for T-44 and by 7% for T-54 compared to T-22.

Based on these results, we are confident that the phonetic information included in the DIMEx100 Corpus is useful for the construction of speech recognition systems, and can be used as seed data to train language technology applications more generally.

8. Conclusions

In this paper we have presented the DIMEx100 corpus as a resource for computational phonetic studies of Mexican Spanish with applications for language technologies. As far as we are aware, this is the largest available empirical resource of this kind, and also the most detailed analysis of phonetic information for this dialect of Spanish. This can be assessed in terms of the number of phonetic units manually tagged by expert-human phoneticians in three different granularity levels of transcription, and also in the number of lexical entries and pronunciations in the pronunciations dictionaries, all of which were identified directly from the corpus.

The design and collection of the corpus responded to the need for a sizable and reliable phonetic resource available for phonetic studies as well as for the construction of acoustic models and pronunciation dictionaries based on direct empirical data. The availability of the Mexbet alphabet and its associated phonetic rules made this effort possible, as before the definition of this alphabet, the set of allophonic units of Mexican Spanish, useful for language technologies, had not been properly identified, and there was confusion on notations and tagging conventions.

We computed the corpus statistics and compared the phonetic distribution with alternative counts for other dialects of Spanish, and the figures suggest that the distribution of samples in the DIMEx100 Corpus reflect the frequency of phonetic units of the language very reasonably. We also used the Corpus to verify a set of phonetic rules describing the expected contexts for this dialect, and compute their corresponding frequency, as shown Table 6. We thus confirmed that most expected contexts do occur in the corpus.

We studied the extent to which the corpus is phonetically complete and balanced. Although we used a measure of perplexity at the level of words for the definition of the corpus, and measured the phonetic figures over the final manual transcription, we verified that there is a good representation of all phonetic units at the three granularity levels. We counted the number of types and instances of diphones and triphones for different amounts of data (i.e. 25%, 50%, 75% and 100%) for all three transcription levels, and identified that the number of types increases very slowly with the amount of data, which suggests that there are very few types in the language that are not

included in the corpus, and these should have very low frequencies. We also identified that the number of high-frequency types is very stable for the four portions of the corpus considered and also for the three levels of transcriptions. From these two observations we conclude that the corpus is reasonably complete and balanced phonetically.

Finally, we validated the corpus as a resource for language technology applications, as was discussed in Section 7. In particular, we tested the quality of the phonetic information contained in the corpus for the construction of acoustic models and pronunciation dictionaries for word recognition at the three levels of transcription, and show that recognizers with different granularity levels can be constructed, with similar recognition rates. We found that the use of finer phonetic transcriptions has a very limited impact on recognition time, in spite of the increased acoustic model size. We hope that the availability of this rich empirical data can be used for further phonetic studies and the construction of language technology applications. In particular, we think that corpus and the present study can be used for training transcription rules for the construction of phonetizers with allophonic variation, with applications in the automatic construction of phonetic dictionaries, and for the automatic tagging of large amounts of speech for more general speaker independent ASR systems. More generally, we think that the present resource can be used as seed-data for training diverse language technology applications for Mexican Spanish.

Acknowledgements

The corpus DIMEx100 has been developed within the context of the DIME Project, at IIMAS, UNAM, with the collaboration of the *Facultad de Filosofía y Letras*, UNAM, and INAOE in Tonanzintla, Puebla. The authors wish to thank the enthusiastic participation of all members of the project who were involved in the collection and transcription of the corpus: Fernanda López, Varinia Estrada, Sergio Coria, Iván Moreno, Ivonne López, Arturo Wong, Laura Pérez, René López, Alejandro Acosta, Alejandro Carrasco, Rafael Torres, Gerardo Mendoza, Ana Ceballos, Alejandra Espinosa and Isabel López; special thanks go to Alejandro Reyes for technical support at INAOE, and to the 100 speakers that provided their voice for the corpus. We also thank James Allen for his continuous collaboration and encouragement along the development of this project. The authors also acknowledge the support of CONACyT's grant 39380-U and PAPIIT-UNAM grant IN121206.

References

Alarcos, E. (1950/1965). **Fonología española**. (Madrid: Gredos)

Canfield, D. L. (1981/1992). **Spanish pronunciation in the Americas**, (Chicago: The University of Chicago Press).

Cuétara, J. (2004). **Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla**. MSc. Dissertation, Universidad Nacional Autónoma de México, México. (In Spanish).

Fetter, P. (1998). **Detection and Transcription of Out-Of-Vocabulary Words in Continuous-Speech Recognition**, PhD thesis, Daimler-Benz AG, aug 1998. Verbmobil Report 231.

Guirao, M. & Borzone, A.M. (1972). **Fonemas, sílabas y palabras en el español de Buenos Aires**, *Filología,* 16, 135-165

Hieronymus, J. L. (1997). **Worldbet phonetic symbols for multilanguage speech recognition and synthesis**. (New Jersey: AT&T and Bell Labs).

Jurafsky, D., Martin, J. H. (2000). **Speech and Language Processing**, (New Jersey: Prentice Hall).

Kirschning, I. (2001) **Research and Development of Speech Technology and Applications for Mexican Spanish at the Tlatoa Group** (Development Consortium at CHI 2001, Seattle, WA.)

Lander, T. (1997). **The CSLU labeling guide**. Oregon: Oregon Graduate Institute of Science and Technology. http://cslu.cse.ogi.edu/corpora/docs/labeling.pdf.

Lope Blanch, J. M. (1963-1964/1983). **En torno a las vocales caedizas del español mexicano,** in *Estudios sobre el español de México*, (pp. 57-77). (México: Universidad Nacional Autónoma de México)

Llisterri, J. & Mariño, J. B. (1993). **Spanish adaptation of SAMPA and automatic phonetic transcription**. Technical Report. SAM-A/UPC/001/v1 -- ESPRIT PROJECT 6819 (SAM-A) *Speech Technology Assessment in Multilingual Applications.* http://liceu.uab.es/~joaquim/publicacions/SAMPA_Spanish_93.pdf

Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M. & Ríos, A. (2003). **The perception of lexical stress in Spanish**, in *Proceedings of the 15th International Congress of Phonetic Sciences.* Barcelona, 3-9 August 2003. pp. 2023-2026. http://liceu.uab.es/~joaquim/publicacions/Llisterri_Machuca_Mota_Riera_Rios_03_Perception_Stress_Spanish.pdf

Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M. & Ríos, A. (2005). **Corpus orales para el desarrollo de las tecnologías del habla en español**. *Oralia. Análisis del discurso oral,* 8, 289-325
http://liceu.uab.es/~joaquim/publicacions/Llisterri_Machuca_Mota_Riera_Rios_05_Corpus_Orales_Tecnologias_Habla_Espanol.pdf

Moreno A., R. Comeyne, K. Haslam, H van den Heuvel, H. Höge, S. Horbach, G. Micca. (2000). **SALA: Speechdat Across Latin America. Results Of The First Phase**, *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Moreno, A., Mariño, J. (1998). **Spanish dialects: Phonetic transcription**, *Proceedings of ICSLP'98, The Fifth Internacional Conference on Spoken Language Processing*, Rundle, Mall: Causal Productions.

Moreno de Alba, J. (1994). **La Pronunciación del Español de México**, El Colegio de México, México.

Navarro Tomás, T. (1918/1970). **Manual de pronunciación española**. Madrid: Consejo Superior de Investigaciones Científicas.

Navarro Tomás, T. (1946/1966). **Escala de frecuencia de fonemas españoles** in *Estudios de fonología española* (pp. 15-30). New York: Las Américas Publishing Company).

NIST (2007). **Speech Recognition Scoring Toolkit (SCTK) Version 2.2.4.** *http://www.nist.gov/speech/tools*.

Pérez, E. H. (2003). **Frecuencia de fonemas**. *e-rthabla, Revista electrónica de Tecnología del Habla* 1. http://lorien.die.upm.es/~lapiz/e-rthabla/numeros/N1/N1_A4.pdf

Perissinotto, G. (1975). **Fonología del español hablado en la Ciudad de México. Ensayo de un método sociolingüístico**. (México: El Colegio de México.)

Pineda, L. A., Massé, A., Meza, I., Salas, M., Schwarz, E., Uraga, E. and Villaseñor, L. (2002). **The DIME Project**, *Proceedings of MICAI2002*, *Lectures Notes in Artificial Intelligence*, Springer-Verlag ,Vol. 2313, 166–175

Pineda, L. A., Villaseñor, L., Cuétara, J., Castellanos, H., López, I. (2004). **DIMEx100: A new phonetic and speech corpus for Mexican Spanish**, en Advances in Artificial Intelligence, Iberamia-2004, C. Lemaitre, C. A Reyes & J. A. Gonzalez (Eds.), *Lectures Notes in Artificial Intelligence,* Springer-Verlag, Vol. 3315, pp. 974–983.

Quilis, A. (1981/1988). **Fonética Acústica de la Lengua Española**. (Madrid: Gredos)

Quilis, A. & Esgueva, M. (1980). **Frecuencia de fonemas en el español hablado**. *Lingüística Española Actual* 2, 1, 1-25.

Ríos Mestre, A. (1999). **La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico del léxico**, *Estudios de Lingüística Española*, Vol.4. *http://elies.rediris.es/elies4/*

Rojo, G. (1991) **Frecuencia de fonemas en español actual**. (In M. Brea and F.M. Fernández Rei (Eds.) *Homenaxe ó profesor Constantino García* (pp. 451-467). Santiago de Compostela: Universidade de Santiago de Compostela, Servicio de Publicación e Intercambio Científico).

Sphinx (2006). **The CMU Sphinx Open Source Speech Recognition Engines.** *http://cmusphinx.sourceforge.net/html/cmusphinx.php*

Strik, H. & Cucchiarini, C. (1998). **Modeling pronunciation variation for ASR: overview and comparison of methods.** In H. Strik, J.M. Kessens, M. Wester (eds.), *Proc. of the ESCA workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, Rolduc, Kerkrade, 4-6 May 1998, pp. 137-144.

Sutton, S., Cole, R., et al. (1998). **Universal Speech Tools: the CSLU Toolkit**. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), pp 3221-3224, Sydney, Australia, November 1998. *http://www.cslu.ogi.edu*.

Wells, J. (1998). **SAMPA. Computer readable phonetic alphabet**. University College London, *http://www.phon.ucl.ac.uk/home/sampa*.

Villaseñor, L., Massé, A. & Pineda, L. (2000). **The DIME Corpus**, Memorias 3º. *Proceedings of Encuentro Internacional de Ciencias de la Computación ENC01*, Tomo II, C. Zozaya, M. Mejía, P. Noriega y A. Sánchez (eds.), SMCC, Aguascalientes, Ags. México, September, 2001.

Villaseñor, L., Montes & Gómez, M., Vaufreydaz, D. & Serignat, J. F. (2004). **Experiments on the Construction of a Phonetically Balanced Corpus from the WEB**, *Proceedings of CICLING2004, LNCS*, Springer-Verlag, Vol. 2945, 416-419.

Appendix 1

Transcription Level T-54

Consonants	**Labial**	**Labio-dental**	**Dental**	**Alveolar**	**Palatal**	**Velar**
Unvoiced stops	[p] ([p_c])		[t] ([t_c])		[k_ j]	[k] ([k_c])
Voiced stops	[b] ([b_c])		[d] ([d_c])			[g] ([g_c])
Unvoiced affricate					[tS] ([tS_c])	
Voiced affricate					[dZ] ([dZ_c])	
Unvoiced fricatives		[f]	[s_ [ ]	[s]		[x]
Voiced fricatives				[z]	[Z]	
Aproximants	[V]		[D]			[G]
Nasals	[m]		[n_ [ ]	[n]	[n~]	[N]
Lateral				[l]		
Tap				[r( ]		
Trill				[r]		

Vowels (unstressed)	**Palatal**				**Central**	**Velar**				
Semi-vowels / semi-consonants	[j]									[w]
Close		[i]							[u]	
Close-mid			[e]					[o]		
Open-mid				[E]			[O]			
Open				[a_ j]	[a]	[a_2]				

Vowels (stressed)	**Palatal**			**Central**				**Velar**
Close	[i_7]							[u_7]
Mid		[e_7]					[o_7]	
			[E_7]			[O_7]		
Open			[a_ j_7]	[a_7]	[a_2_7]			

Appendix 2

Transcription Level T-44

Consonants	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar
Unvoiced stops	[p] ([p_c])		[t] ([t_c])			[k] ([k_c])
Voiced stops	[b] ([b_c])		[d] ([d_c])			[g] ([g_c])
Unvoiced affricate					[tS] ([tS_c])	
Unvoiced fricatives		[f]		[s]		[x]
Voiced fricatives					[Z]	
Aproximants	[V]		[D]			[G]
Nasals	[m]			[n]	[n~]	
Lateral				[l]		
Tap				[r( ]		
Trill				[r]		

Vowels (unstressed)	Palatal			Central	Velar		
Semi-vowels / semi-consonants	[j]						[w]
Close		[i]				[u]	
Mid			[e]		[o]		
Open				[a]			

Vowels (stressed)	Palatal		Central	Velar	
Close	[i_7]				[u_7]
Mid		[e_7]		[o_7]	
Open			[a_7]		

	Syllable coda
Labial /p – b/	[-B]
Dental /t – d/	[-D]
Velar /k – g/	[-G]
Nasals /n – m/	[-N]
Trill and Tap /r(- r/	[-R]

Appendix 3

Transcription Level T-22

Consonants	**Labial**	**Labio-Dental**	**Dental**	**Alveolar**	**Palatal**	**Velar**
Unvoiced stops	[p]		[t]			[k]
Voiced stops	[b]		[d]			[g]
Unvoiced affricate					[tS]	
Unvoiced fricatives		[f]		[s]		[x]
Voiced fricatives					[Z]	
Nasals	[m]			[n]	[n~]	
Lateral				[l]		
Tap				[r( ]		
Trill				[r]		

Vowels	**Palatal**		**Central**	**Velar**	
Close	[i]				[u]
Mid		[e]		[o]	
Open			[a]		

Appendix 4

Mean time duration of phonetic units (in miliseconds) in the levels T54, T44 and T22

Units	Samples	Mean	Std. Dev.
[p_c]	6,730	66.28	23.36
[p]	6,730	19.51	22.90
[t_c]	12,242	54.04	18.96
[t]	12,246	23.19	22.85
[k_c]	9,748	53.59	19.69
[k]	8,464	27.65	10.83
[k_j]	1,285	30.75	10.87
[b_c]	1,229	33.30	28.59
[b]	1,303	22.66	22.45
[V]	4,186	53.38	22.66
[d_c]	3,699	30.63	17.98
[d]	3,881	22.36	13.36
[D]	10,115	47.00	32.44
[g_c]	421	30.82	18.36
[g]	426	27.94	14.56
[G]	1,899	56.44	27.83
[tS_c]	386	50.20	16.46
[tS]	385	64.62	23.58
[f]	2,116	87.56	22.96
[s]	20,926	95.28	19.04
[s_[]	2,912	61.90	9.89
[z]	2,123	53.29	14.59
[x]	1,994	93.46	22.33
[Z]	720	76.98	16.08
[dZ_c]	127	43.62	17.64
[dZ]	126	43.27	18.49
[m]	7,718	74.41	17.75

Units	Samples	Mean	Std. Dev.
[n]	12,021	65.50	25.88
[n_[]	4,899	65.89	16.68
[N]	848	63.39	28.78
[n~]	346	86.68	32.69
[l]	14,058	64.08	24.28
[r(]	14,784	45.29	36.51
[r]	1,625	76.50	20.80
[i]	9,705	59.15	20.53
[i_7]	3,941	80.57	27.84
[j]	8,349	52.68	23.12
[e]	23,434	61.72	24.76
[e_7]	6,883	73.57	27.53
[E]	3,083	62.89	25.83
[E_7]	1,153	84.12	24.58
[a]	18,927	75.69	22.16
[a_7]	8,022	89.10	30.76
[a_j]	539	72.65	15.35
[a_j_7]	228	95.14	19.54
[a_2]	1,277	66.73	17.57
[a_2_7]	1,164	85.58	23.58
[o]	15,088	67.58	28.47
[o_7]	4,200	71.90	7.62
[O]	3,064	63.30	28.49
[O_7]	1,533	76.12	17.70
[u]	3,431	56.31	24.82
[u_7]	1,716	75.20	21.40
[w]	2,752	49.45	18.49

Level T-54

Units	Samples	Mean	Std. Dev.
[p_c]	6,573	66.53	22.86
[p]	6,571	19.47	7.40
[t_c]	12,115	53.84	22.61
[t]	12,117	22.95	8.86
[k_c]	8,437	56.11	19.15
[k]	8,440	28.71	10.18
[b_c]	1,138	31.50	21.01
[b]	1,213	22.19	16.99
[-B]	287	78.48	33.94
[V]	4,141	53.18	12.88
[d_c]	3,518	29.50	21.18
[d]	3,707	21.80	14.99
[D]	9,663	45.64	13.63
[-D]	735	86.46	48.98
[g_c]	328	30.26	19.17
[g]	334	28.04	15.64
[G]	1,745	56.49	14.39
[-G]	1,548	60.03	20.32
[tS_c]	385	50.13	16.43
[tS]	384	64.51	23.56
[f]	2,111	87.48	22.90
[s]	25,920	88.12	36.74

Units	Samples	Mean	Std. Dev.
[x]	1,991	93.48	22.32
[Z]	841	86.32	112.43
[m]	6,076	75.93	16.61
[n]	7,920	65.53	17.00
[-N]	11,471	65.96	26.65
[n~]	346	86.68	18.49
[l]	14,049	64.06	24.27
[r(]	10,016	39.38	11.33
[r]	1,607	76.37	23.19
[-R]	4,767	57.86	29.18
[i]	9,694	59.14	20.52
[i_7]	3,936	80.56	27.85
[j]	8,337	52.69	23.10
[e]	26,496	61.85	24.64
[e_7]	8,030	75.09	28.24
[a]	20,734	75.07	32.05
[a_7]	9,402	88.81	27.52
[o]	18,136	66.86	31.24
[o_7]	5,724	73.02	28.55
[u]	3,436	56.40	19.13
[u_7]	1,718	75.20	24.81
[w]	2,744	49.31	21.17

Level T-44

Units	Samples	Mean	Std. Dev.
[p]	6683	86.18	30.71
[t]	12152	77.43	25.68
[k]	9661	81.59	38.82
[b]	5431	53.59	19.00
[d]	13851	48.76	27.25
[g]	2297	57.20	21.91
[tS]	382	115.29	33.52
[f]	2100	87.49	22.98
[s]	25739	88.04	36.77
[x]	1979	93.51	22.37
[Z]	836	76.07	38.36
[m]	7647	74.43	17.82
[n]	17629	65.52	23.46
[n~]	341	86.57	18.36
[l]	13934	64.10	24.33
[r(]	14654	45.32	20.82
[r]	1609	76.46	23.37
[i]	21772	60.54	24.99
[e]	34236	64.96	26.18
[a]	29893	79.38	31.42
[o]	23682	68.36	30.81
[u]	7825	58.04	23.25

Level T-22

Appendix 5

Equivalent symbols between IPA and Mexbet

Consonants	**IPA**	**Mexbet**
Labial unvoiced stop	p	p
Dental unvoiced stop	t	t
Velar unvoiced stop	k	k
Palatalized unvoiced stop	k^j	k_j
Labial voiced stop	b	b
Dental voiced stop	d	d
Velar voiced stop	g	g
Palatal unvoiced affricate	$\widehat{t\int}$	tS
Palatal voiced affricate	$\widehat{d\textyogh}$	dZ
Labiodental unvoiced fricative	f	f
Alveolar unvoiced fricative	s	s
Dentalized unvoiced fricative	s̪	s_[
Velar unvoiced fricative	x	x
Alveolar voiced fricative	s̬	z
Palatal voiced fricative	ʒ	Z
Labial approximant	β̞	V
Dental approximant	ð̞	D
Velar approximant	ɣ̞	G
Labial nasal	m	m
Dentalized nasal	n̪	n_[
Alveolar nasal	n	n
Palatal nasal	ɲ	n~
Velarized nasal	$n^ɤ$	N
Lateral	l	l
Tap	ɾ	r(
Trill	r	r

Vowels	**IPA**	**Mexbet**
Palatal semi-vowel/consonant	i̯ / j	j
Close palatal	i	i
Mid palatal	e	e
Mid palatal oponed	e̞	E
Palatalized central open	a^{+}	a_j
Central open	a	a
Velarized central open	a̠	a_2
Mid velar opened	o̞	O
Open velar	o	o
Close velar	u	u
Velar semi-vowel/consonant	u̯ / w	w