DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish

Luis A. Pineda[a], Luis Villaseñor Pineda[b], Javier Cuétara[c],
Hayde Castellanos[a,c], and Ivonne López[a,c]

[a] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS, UNAM)
luis@leibniz.iimas.unam.mx,
{haydecv, ivlom}@turing.iimas.unam.mx
[b] Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE)
villasen@inaoep.mx
[c] Facultad de Filosofía y Letras, UNAM
j_cuetara@correo.unam.mx

Abstract. In this paper the phonetic and speech corpus DIMEx100 for Mexican Spanish is presented. We discuss both the linguistic motivation and the computational tools employed for the design, collection and transcription of the corpus. The phonetic transcription methodology is based on recent empirical studies proposing a new basic set of allophones and phonological rules for the dialect of the central part of Mexico. These phonological rules have been implemented in a visualization tool that provides the expected phonetic representation of a text, and also a default temporal alignment between the spoken corpus and its phonetic representation. The tools are also used to compute the properties of the corpus and compare these figures with previous work.

1 Introduction

Despite recent progress in speech recognition, the availability of phonetic corpora for the creation of acoustic models in Spanish is still very limited[1]. The creation of this kind of resources is required for a variety of reasons: TTSs (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. A linguistic and empirically motivated allophonic set is also important for the definition of pronunciation dictionaries. Mexican Spanish, for instance, is a language with 22 phones (17 consonants and 5 vowels), 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

[1] Among the few available corpus we can mention the Latino-40 Spanish Read News for Latin-American Spanish, available LDC (http://www.ldc.upenn.edu), the ALBAYZIN Corpus for Castillan Spanish and the Spanish Speech Corpus 1, these later two available at ELDA (http://www.elda.fr).

C. Lemaître, C.A. Reyes, and J.A. Gonzalez (Eds.): IBERAMIA 2004, LNAI 3315, pp. 974–983, 2004.

often enough in spoken language to be considered in transcriptions and phonetic dictionaries, and whose acoustic properties deserve the definition of an acoustic model for each unit in this set. Previous corpora for Mexican Spanish, like Tlatoa (Kirshning, 2001) and Gamboa (2002) 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 sounds (e.g. [j] and [w]). Another consideration is that phonetic contexts depend on the dialect, and it is possible to define a set of phonological rules that predict the phonetic units that are likely to be pronounced by a speaker of the dialect in a given context. This information is useful not only for the TTS, but also for the definition of automatic tools that help the transcription process.

In this paper we present the design and collection process of the DIMEx100 corpus, including its linguistic and empirical motivation and the computational methodology and tools employed and developed for its transcription. We also present an assessment of the corpus properties in relation to previous work.

2 Corpus Design and Characteristics

For the collection process we considered the Web as a large enough, complete and balanced, linguistic resource, and selected the corpus phrases from this source; the result of this exercise was the Corpus230 (Villaseñor *et al.*, 2004) consisting of 344,619 phrases with 235,891 lexical units, and about 15 million words. From this original resource we selected 15,000 phrases from 5 to 15 words; these phrases were ordered according to its perplexity value from lowest to highest; intuitively the perplexity is a measure of the number of linguistic units that can follow a reference unit in relation to the corpus: the lower the perplexity of a word, for instance, the less the number of different words that are likely to follow it in a sentence; the perplexity of a sentence can then be defined as a function of the perplexity of its constituent words; accordingly, sentences with a low perplexity are constituted by words with a high discriminating power or information content in relation to the corpus. For the final corpus we took the 7000 sentences with the lowest perplexity value. Phrases 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). Finally we arrived at a set of 5010 phrases; the corpus was recorded by 100 speakers, each recorded 50 individual phrases, in addition to 10 phrases that were recorded by all 100 speakers. With this we arrived to 6000 phrases: 5000 different phrases recorded one time and 10 phrases recorded 100 times each. The final resource is the DIMEx100 corpus. In order to measure the appropriateness of the corpus we controlled the characteristics of the speakers; we also measured the amount and 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 study at CCADET, UNAM, with a Single Diaphragm Studio Condenser Microphone Behringe B-1, a Sound Blaster Audigy

Platimun ex (24 bit/96khz/100db SNR) using the Wave Labe program; the sampling format is mono at 16 bits, and the sampling rate is 44.1 khz.

2.1 Socio-linguistic Characteristics of the Speakers

Recording an oral corpus implies considering and designing minimal linguistic measurable aspects in order to be able to evaluate them afterwards. Following Perissinotto (1975), the 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, teachers and workers) 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, 18 people from other places residing in Mexico City participated in the recordings. The corpus resulted gender balanced (49% males; 51% 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).

2.2 Linguistic Characteristics

From our empirical study (Cuétara, 2004) of the DIME Corpus (Villaseñor *et al.*, 2000), developed within the context of the DIME Project (Pineda *et al.*, 2002), a set of 37 allophones for Mexican Spanish that occur often enough and can be clearly distinguished was identified; this set is being used for the transcription process. This set is illustrated in Table 1. The Mexbet phonetic alphabet (Uraga and Pineda, 2002; Cuétara, 2004) is used for the representation of the phonetic units, but the table should be self explanatory given the articulation point for each allophone and the distinctive features in the left column.

In addition to this allophonic set we have identified the context in which all of these allophones occur. These contexts have been characterized through phonological rules; we present two instances for illustration: the rule for the bilabial voiced stop /b/ and the rule for the open vowel /a/. In table 2 "///_" stands for an initial context; "b_c" stands for the closure of the "b" allophone (which is considered as a unit in phonetic transcriptions), and "V" stands for the fricative realization of /b/. So, the unvoiced stop /b/ occurs in initial contexts, and after /m/ or /n/, and its fricative realization everywhere else. The vowel /a/, in turn, can be realized as a velar in front of other velar sounds and in front of /l/ in the coda of a syllable (e.g. *alto*); this is indicated by the "$" sign in "_{l}$". The palatal /a/, in turn, is realized in front of palatal sounds, and the central /a/ elsewhere. In general, the allophonic variation of Spanish can be modeled with phonological rules (Moreno and Mariño, 1998), and the full set applicable to Mexican Spanish is presented in (Cuétara, 2004).

Table 1. Allophonic set for Mexican Spanish

<table>
<tr><th colspan="2"></th><th>Labial</th><th>Labio-dental</th><th>Dental</th><th>Alveolar</th><th colspan="6">Palatal</th><th colspan="4">Velar</th></tr>
<tr><td colspan="2">Unvoiced stops</td><td>p</td><td></td><td>t</td><td></td><td colspan="6">k_ j
<i>“queso”, “kilo”</i></td><td colspan="4">k</td></tr>
<tr><td colspan="2">Voiced stops</td><td>b</td><td></td><td>d</td><td></td><td colspan="6"></td><td colspan="4">g</td></tr>
<tr><td colspan="2">Unvoiced affricate</td><td></td><td></td><td></td><td></td><td colspan="6">tS
<i>“hacha”</i></td><td colspan="4"></td></tr>
<tr><td colspan="2">Voiced africate</td><td></td><td></td><td></td><td></td><td colspan="6">dZ
<i>“lluvia”, “un yunque”</i></td><td colspan="4"></td></tr>
<tr><td colspan="2">Unvoiced fricatives</td><td></td><td>f</td><td>s_ [
<i>“asta”</i></td><td>S</td><td colspan="6"></td><td colspan="4">x</td></tr>
<tr><td colspan="2">Voiced fricatives</td><td>V
<i>“haba”</i></td><td></td><td>D
<i>“hada”</i></td><td>z
<i>“mismo”</i></td><td colspan="6">Z
<i>“ayer”, “mi yunque”</i></td><td colspan="4">G
<i>“haga”</i></td></tr>
<tr><td colspan="2">Nasals</td><td>m</td><td></td><td>n_ [
<i>“antes”</i></td><td>n</td><td colspan="6">n~
<i>“año”</i></td><td colspan="4">N
<i>“angel”</i></td></tr>
<tr><td rowspan="2">Liquids</td><td>Lateral</td><td></td><td></td><td></td><td>l</td><td colspan="6"></td><td colspan="4"></td></tr>
<tr><td>Vibrants</td><td></td><td></td><td></td><td>r(
<i>“pero”</i>
r
<i>“perro”</i></td><td colspan="6"></td><td colspan="4"></td></tr>
<tr><td rowspan="4">Vowels</td><td>High</td><td></td><td></td><td></td><td></td><td>j
<i>“aire”</i></td><td>i</td><td></td><td></td><td></td><td colspan="2"></td><td></td><td>u</td><td>w
<i>“aura”</i></td></tr>
<tr><td rowspan="2">Medium</td><td></td><td></td><td></td><td></td><td></td><td></td><td>e</td><td></td><td></td><td colspan="2"></td><td>o</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>E
<i>“erre”</i></td><td></td><td colspan="2">O
<i>“sol”</i></td><td></td><td></td><td></td></tr>
<tr><td>Low</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>a_ j
<i>“aire”</i></td><td>a</td><td colspan="2">a_2
<i>“aunque”, “alma”</i></td><td></td><td></td><td></td></tr>
</table>

Table 2. Examples of Phonological Rules

Bilabial voiced stop b		
/b/	b_c b	///_
/b/	b_c b	{m, n}_
/b/	V	Elsewhere

Open vowel a		
/a/	a_2	_{u, x}
/a/	a_2	_{l}$
/a/	a_j	_{tS, n~, Z, j}
/a/	a	Elsewhere

3 Computational Tools

In this section we present three computational tools that we have developed to assist human transcribers in the transcription process:

1. A set of transcription rules mapping textual representations (graphemes) to their corresponding phonological and phonetic representations (i.e. transcriptions).
2. An interactive visualization tool or “phonetizer” that translates a given text into its phonological and phonetic representation using these grapheme to phoneme and grapheme to allophone conversion rules.

3. A script that given a speech signal and its associated textual transcription produces the expected segmental representation with a default temporal alignment between the phonetic and allophonic units and its associated speech signal.

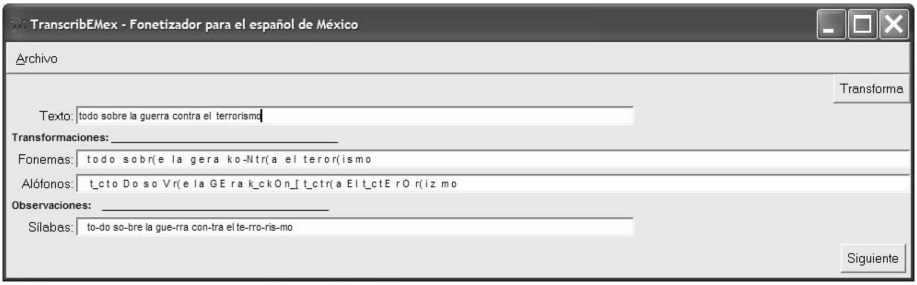
3.1 Textual to Phonetic Representations Translation Rules

For this translation, the textual representation is first divided into syllables. A syllabizer that follows the standard syllabic structure of Spanish has been implemented with standard regular expressions in Perl. The syllabic representation is then read left to right scanning the orthographic symbols one at the time in two stages: first, orthographic symbols are substituted by their corresponding phoneme representation; in Spanish, the mapping from graphemes to phonemes is not one to one: a given symbol can have no phonetic realization (e.g. *h* is never realized, or *u* in the context of *g_e*), and more than one (e.g. *c* as [k] and [s])[2]; also, two symbols can be realized as one sound (e.g. *ch* as [tS], *y* and *ll* as [Z] or [dZ]), etc. We have developed set of rules including one for each possible situations. Orthographic contexts are represented through regular expressions, and substitution rules are applied from the most specific to the most general context. In the second stage, phoneme representations are substituted by their corresponding allophonic representations; for this translation, a regular expression matching the context for every symbol, as shown in Table 2, is defined. Rules are applied from the most specific to the most general context too; for instance, for the orthographic translation of /b/ the rule for initial context is tried first, next the rule for the nasal left context and finally the default context rule that rewrites the fricative realization; similarly for the open vowel *a*. Archiphonemes are also considered: b/p, d/t, g/k, n/m and r(/r have no contrasting value when they appear in the coda of a syllable and both of the elements in each pair are subsumed in the common representation /-B/, /-D/, /-G/, /-N/ and /-R/ respectively (e.g. *optar* is transcribed as /o-Bta-R/ and *camión* as /kamio-N/).

3.2 The Phonetizer

The grapheme to phoneme and allophonic representations were used to build a visualization tool, using Perl and tcl/Tk. With this interface the user can type the text, and the phonemic and allophonic transcriptions are displayed. The tool provides an initial expectation for the human transcriber in the temporal alignment process. It is also useful for didactical purposes and has been used to train human transcribers in the DIME-II Project, and has been received very positively by undergraduate students of phonetic courses at UNAM too. The phonetizer is illustrated in Figure 1.

[2] The *x* grapheme has at least three pronunciations: [ks], [x] and [s]. Currently, we give the standard pronunciation [ks], but we plan to extend the rules with a pronunciation dictionary for Mexican Spanish. We also have cases of grapheme to grapheme conversion: *w* is conventionally rewritten as *gü*.



Texto: todo sobre la guerra contra el terrorismo

Transformaciones:

Fonemas: todo sobr(e la gera ko-Ntr(a el teror(ismo

Alófonos: t_cto Do so Vr(e la GE ra k_ckOn_[t_ctr(a El t_ctE rO r(iz mo

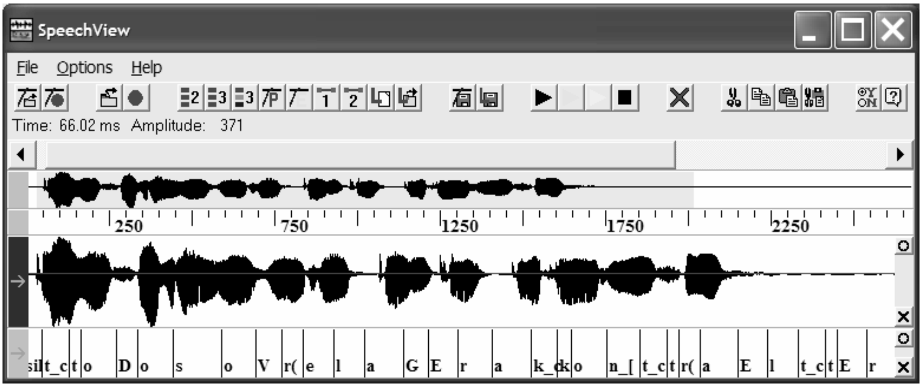
Observaciones:

Sílabas: to-do so-bre la gue-rra con-tra el te-rro-ris-mo

Fig. 1. Phonetic representations visualization tools

3.3 Default Transcription with Temporal Alignment

The transcription rules have also been used to construct a tool that given a speech signal and its orthographic transcription produces a default phonetic transcription temporally aligned with the speech data; mean temporal durations were computed for all allophones from the DIME Corpus, and this figures are continuously updated with the data produced by the human transcription of DIMEx100 itself. We are using the CSLU Tool Kit´s SpeechView[3] program for the human transcription, and the default temporal alignment is produced in the format for tags of this program (.phn). This tool facilitates greatly the phonetic transcription task: the human transcriber simple loads the speech and the tag´s files, and completes the transcription manually. An illustration of the SpeechView interface, with the speech data and the default tags is shown in Figure 2.



SpeechView
File Options Help
Time: 66.02 ms Amplitude: 371
250
750
1250
1750
2250
sil t_c t o D o s o V r(e l a G E r a k_c k o n_[t_c t r(a E l t_c t E r

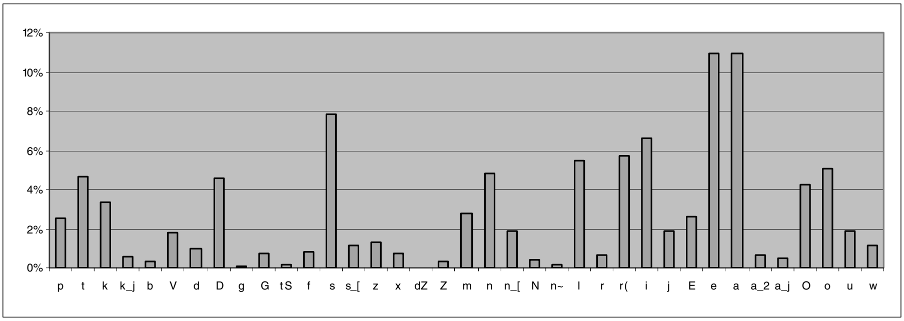
Fig. 2. Default temporal alignments

[3] http://cslu.cse.ogi.edu

4 Corpus Statistics and Evaluation

The text to phonetic representation rules also allow us to evaluate whether the corpus is complete, large enough and balanced. For this we simply translate the text into its phonemic and allophonic representations, and compute the number and distribution of samples; as expected, the corpus includes all phonetics units, with a large enough number of samples for the training process. In particular, the less represented phonetic units are [n~] with 338 samples, [g] with 227 and [dZ] with 23; with the exception of this latter allophone, which can be subsumed with [Z] or complemented with supplementary sample words, the figures are satisfactory. Table 3 shows the expected distribution of the corpus according to rules for all 5010 different phrases; the actual figures for the spoken phrases (i.e. the speech corpus) will be reported when the human transcription is available. Also, the set of phonological rules for all allophones specifies all possible contexts in which these units can occur; the contexts in the right column of Table 2 for all the allophones abstract over all contexts of the form $\alpha\beta\gamma$, where α and γ are the left and right allophones in relation to a reference allophone β. As we have a significant number of instances of all allophones in the corpus, all the contexts specified in the phonological rules appear systematically, and we conclude that the corpus is complete[4]. 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.

Table 3. Phonetic distribution of DIMEx100

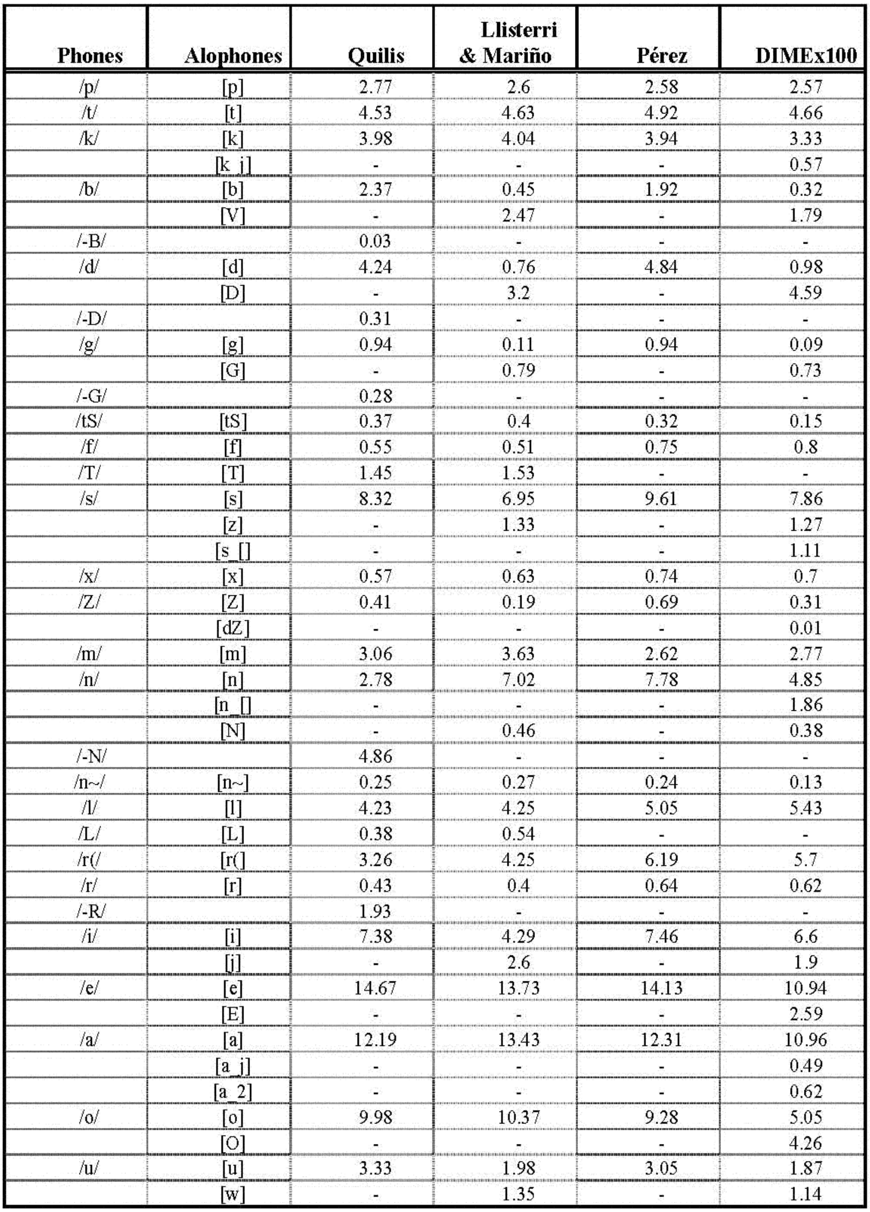


These figures can also be used to assess whether the corpus is balanced. In Table 4 we compare the distribution of DIMEx100 in relation to Quillis (1981), Llisterri and Mariño (1993) for peninsular Spanish and Pérez (2003) for Chilean Spanish. As can be seen in Table 4 our balancing procedure follows closely the figures of previous studies, taken into account the allophonic differences between the dialects, and we

[4] For an alternative balancing method see (Uraga and Gamboa, 2004).

conclude that DIMEx100 is balanced. The correlation at the level of the phonemes between DIMEx100 and these three corpora is shown in Table 5.

Table 4. Phonetic distribution of Spanish



Phones	Alophones	Quilis	Llisterri & Mariño	Pérez	DIMEx100
/p/	[p]	2.77	2.6	2.58	2.57
/t/	[t]	4.53	4.63	4.92	4.66
/k/	[k]	3.98	4.04	3.94	3.33
	[k_j]	-	-	-	0.57
/b/	[b]	2.37	0.45	1.92	0.32
	[V]	-	2.47	-	1.79
/-B/		0.03	-	-	-
/d/	[d]	4.24	0.76	4.84	0.98
	[D]	-	3.2	-	4.59
/-D/		0.31	-	-	-
/g/	[g]	0.94	0.11	0.94	0.09
	[G]	-	0.79	-	0.73
/-G/		0.28	-	-	-
/tS/	[tS]	0.37	0.4	0.32	0.15
/f/	[f]	0.55	0.51	0.75	0.8
/T/	[T]	1.45	1.53	-	-
/s/	[s]	8.32	6.95	9.61	7.86
	[z]	-	1.33	-	1.27
	[s_[]	-	-	-	1.11
/x/	[x]	0.57	0.63	0.74	0.7
/Z/	[Z]	0.41	0.19	0.69	0.31
	[dZ]	-	-	-	0.01
/m/	[m]	3.06	3.63	2.62	2.77
/n/	[n]	2.78	7.02	7.78	4.85
	[n_[]	-	-	-	1.86
	[N]	-	0.46	-	0.38
/-N/		4.86	-	-	-
/n~/	[n~]	0.25	0.27	0.24	0.13
/l/	[l]	4.23	4.25	5.05	5.43
/L/	[L]	0.38	0.54	-	-
/r(/	[r(]	3.26	4.25	6.19	5.7
/r/	[r]	0.43	0.4	0.64	0.62
/-R/		1.93	-	-	-
/i/	[i]	7.38	4.29	7.46	6.6
	[j]	-	2.6	-	1.9
/e/	[e]	14.67	13.73	14.13	10.94
	[E]	-	-	-	2.59
/a/	[a]	12.19	13.43	12.31	10.96
	[a_j]	-	-	-	0.49
	[a_2]	-	-	-	0.62
/o/	[o]	9.98	10.37	9.28	5.05
	[O]	-	-	-	4.26
/u/	[u]	3.33	1.98	3.05	1.87
	[w]	-	1.35	-	1.14

Table 5. Correlation of DIMEx100 with previous corpora

Coeficientes de correlación contra DIMEx100	
Quilis	0.97
Llisterri & Mariño	0.98
Pérez	0.99

5 Conclusions

In this paper we have presented the DIMEx100 corpus for the creation of acoustic models for Mexican Spanish. The design methodology and corpus statistics have also been reported. We have considered the Web as a large enough, complete and balanced linguistic resource and a simple method for the corpus design, based on an entropy measure, was developed; this method was used with very satisfactory results as the resulting corpus is complete and balanced. The transcription, currently under development, is based on a set of allophones and phonological rules that were identified through an empirical investigation of the DIME Corpus. The set of phonetic units and phonological rules has also been used for the development of computational tools to support the transcription process. The main tools are “a phonetizer”, that given a text returns its phonemic and allophonic transcription, and a script that given a speech signal and its orthographic transcription returns an allophonic transcription with a default temporal alignment; the first tool has also been used as a tutorial aid in phonetics and phonology, both for the members of the DIME-II Project, and also in formal phonetics courses in the school of Spanish linguistics at UNAM´s *Facultad de Filosofía y Letras*. The tools have also used to measure the expected number and distribution of phonetic units in the corpus, and we have been able to assess the corpus in relation to previous work, with very positive results.

Acknowledgements

The corpus DIMEx100 has been developed within the context of the DIME-II Project, at IIMAS, UNAM, with the collaboration of *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 participated in the collection and transcription of the corpus (in progress): Fernanda López, Varinia Estrada, Sergio Coria, Patricia Pérez, Iván Moreno and Isabel López; also to the 100 locutors that participated in this effort. We also wish to thank the audio support provided by José Mondragón at CCADET and the computational support provided by César Gamboa at DCC-IIMAS. We also wish to acknowledge the facilities provided by Dr. Felipe Lara and Dr. Fernando Gamboa to record the corpus at CCADET, UNAM. The authors also acknowledge the support of Conacyt´s grant 39380-U and LAFMI (Laboratorio Franco Mexicano de Informática).

References

1. Canfield, D. L. (1981/1992). Spanish pronunciation in the Americas, Chicago: The University of Chicago Press.
2. Cuétara, J. (2004). Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla. MSc. Thesis in Spanish Linguistics, UNAM, Mex (In Spanish).
3. Gamboa, C. (2001). Un sistema de reconocimiento de voz para el Español. BSc. Thesis. UNAM, Mex (In Spanish).
4. 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.
5. Lope Blanch, J. M. 1963-1964/1983. En torno a las vocales caedizas del español mexicano, en Estudios sobre el español de México, México: Universidad Nacional Autónoma de México, pp. 57-77 (In Spanish).
6. Llisterri, J. and Mariño, J. B. (1993). "Spanish adaptation of SAMPA and automatic phonetic transcription". Reporte técnico del ESPRIT PROJECT 6819, Speech Technology Assessment in Multilingual Applications, 9 pp.
7. Moreno, A. and Mariño, J. B. (1998). Spanish dialects: Phonetic transcription, Proceedings of ICSLP'98. The 5th International Conference on Spoken Language Processing, Sydney.
8. Pérez, H. E. (2003). Frecuencia de fonemas, Concepción: Universidad de Concepción, Chile. (In Spanish).
9. Perissinotto, G. 1975. Fonología del español hablado en la Ciudad de México. Ensayo de un método sociolingüístico, México: El Colegio de México (In Spanish).
10. 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, Vol. 2313, Springer-Verlag.
11. Quilis, A. (1981/1988). Fonética Acústica de la Lengua Española, Madrid: Gredos. (In Spanish).
12. Uraga, E. and Pineda, L. A. (2002). Automatic generation of pronunciations lexicons for Spanish, Computational Linguistics and Intelligent Text Processing, Third International Conference CICLing 2002, Alexander Gelbuck (ed.), Lecture Notes in Computer Science, Vol. 2276, Springer-Verlag, pp. 330-339, Berlin.
13. Uraga, E. and Gamboa, C. (2004). VOXMEX Speech Database: design of a phonetically balanced corpus, Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, May 2004.
14. Villaseñor, L., Massé, A. and Pineda, L. (2000). A Multimodal Dialogue Contribution Coding Scheme, Proceedings of ISLE workshop, LREC2000 Athens, May 29-30.
15. Villaseñor, L., Montes y Gómez, M., Vaufreydaz, D. and Serignat, J. F. (2004). Experiments on the Construction of a Phonetically Balanced Corpus from the WEB, Proceedings of CICLING2004, LNCS, Vol. 2945, Springer-Verlag.