Novel Online Tools for Automatic Generation of Pronouncing Dictionaries in Mexican Spanish for Speech Processing

Carlos D. Hernández-Mena and Abel Herrera-Camacho
Signal Processing Department, National Autonomous University of Mexico (UNAM), Mexico City
Email: ca_hernandez@uxmcc2.iimas.unam.mx, Abelherrerac1@gmail.com

***Abstract*—A pronouncing dictionary is a very important tool in a speech processing system. In speech recognition, it helps to the training stage to create the Markov models for every phoneme of every word in the lexicon. In speech synthesis, it helps the system to produce the correct pronunciations of the words introduced by the user despite the orthographic representations of them. All of this implies that creation of pronouncing dictionaries depends on the language one has chosen, because different languages have different sets of phonemes and also different phonetic alphabets to represent them. In this paper we present a methodology of creation of pronouncing dictionaries in Mexican Spanish utilizing a set of novel online tools developed by the CIEMPIESS-UNAM Project. These tools are for free use and they produce pronouncing dictionaries in a particular phonetic alphabet called Mexbet.**

***Index Terms*—Mexican Spanish, Mexbet, pronouncing dictionary, phonetic alphabet**

I. Introduction

As we know, the fields of phonetics and phonology are closely linked to the speech processing. Phonetics is the branch of linguistics that deals with the acoustic realization of the speech sounds and it is interested in the amplitude, frequency, duration and other measurable parameters of the speech signal. Phonology studies the speech sounds as an abstract system. This means that in phonology these speech sounds are represented only in a prototypic way. A main difference between phonology and phonetics is that the former deals with ideal sounds called phonemes and the latter investigates how these phonemes vary from person to person. The variants of the prototypic phonemes are also known as allophones [1], [2]. A phonetic alphabet is a set of symbols that represents phonemes and allophones, and many different languages have their own phonetic alphabets. For example, a phonetic alphabet, exclusive for the Spanish language is the RFE [3]. A more standardized alphabet that is used to represent most of the languages over the world would be the International Phonetic Alphabet [4]. Nevertheless, there is a problem with these "classic" phonetic alphabets in the field of computer science: the symbols they utilize have not a common character codification like ASCII or UTF-8. This means that we can not use them in programming codes and that is why the solution to this, is the creation of ASCII Computational Phonetic Alphabets (ACPA for short). The ACPA are made up only by ASCII symbols and that is why they can be easily incorporated in programming codes. Some examples of ACPA are the "Speech Assessment Methodology Phonetic Alphabet" (SAMPA) [5] that was designed for many languages including Spanish [6], the WORLDBET alphabet that is an ASCII adaptation of the IPA alphabet [7] or the OGIBET that was created by the Oregon Institute of Science and Technology [8], and then adapted for the Mexican Spanish by the Tlatoa Group in Puebla, Mexico [9].

In the field of speech processing one can use an ACPA to produce pronouncing dictionaries. A pronouncing dictionary (PD) is nothing but a list of words followed by their phonetic transcription in a particular ACPA. Fig. 1 shows how a PD written in Mexbet looks like.

```
azotobactEr a s o t o b a k t e_7 r(
azoyU a s o Z u_7
azpIri a s p i_7 r( i
aztEca a s t e_7 k a
aztEcas a s t e_7 k a s
azuAra a s u a_7 r( a
azucEna a s u s e_7 n a
azucarAdas a s u k a r( a_7 d a s
azucarAdo a s u k a r( a_7 d o
```

Figure 1. Pronouncing dictionary in Mexbet.

Mexbet is an example of an ACPA that was specially design for the Spanish spoken at Central Mexico [10] and we (the CIEMPIESS-UNAM Project) chose it to develop our automatic tools.

The CIEMPIESS-UNAM Project aims to develop and share free and open-source tools for speech processing in the Spanish language. It was created at the Speech Processing Laboratory of the Faculty of Engineering of the UNAM (FI-UNAM) and it has recently published an article about a new 17 hours radio corpus called CIEMPIESS [11] that is available for free use at the project website[1].

Table I shows the symbols of Mexbet and their equivalent to the IPA symbols.

Manuscript received August 20, 2014; revised November 21, 2014.

[1] CIEMPIES-UNAM Project website at: www.ciempiess.org

doi: 10.12720/ijsps.4.1.51-54

TABLE I. MEXBET SYMBOLS AND THEIR EQUIVALENT TO IPA SYMBOLS.

IPA	Mexbet	IPA	Mexbet
p	p	n	n
t	t	ɾ	r(
k	k	r	r
b	b	ɲ	n~
d	d	l	l
g	g	f	f
t͡ʃ	tS	a	a
s	s	e	e
ʃ	S	o	o
x	x	i	i
ʝ	Z	u	u
m	m	t͡l	tl

Notice that all the Mexbet symbols are actually made up by ASCII characters.

In the present document we will show a set of techniques to create accurate PD's in Mexbet using the online tools developed by the CIEMPIESS-UNAM Project that are for free use, and available at the project website. At the end, we will also evaluate the results utilizing two different Mexican Spanish corpuses (databases).

II. TOOLS AND METHODOLOGIES

There are three different online tools which help to generate PD's automatically. All of them require putting the input data into a UTF-8 plain text file. Once you select this file with the option "Seleccionar archivo", you have to press a button with the caption "PROCESAR ARCHIVO" to process the file. Then the tool will convert the input data into a PD and show you a link to download the resulting file. All the words in the PD are unique and they are sorted alphabetically.

A. Pronouncing Dictionary from a Raw Text

TABLE II. EXAMPLES OF PREPROCESSED WORDS AND THEN TRANSCRIBED IN MEXBET

Word	Preprocessing	Mexbet
congelado	congelAdo	k o n x e l a_7 d o
alcantarilla	alcantarIlla	a l k a n t a r(i_7 Z a
peñasco	peñAsco	p e n~ a_7 s k o
caza	cAza	k a_7 s a
acción	acciOn	a k s i o_7 n
chamaco	chamAco	tS a m a_7 k o
correo	corrEo	k o r e_7 o
sharon	SAron	S a_7 r(o n
sexenio	sexEnio	s e k s e_7 n i o
xilófono	$ilOfono	s i l o_7 f o n o
xavier	JaviEr	x a b i e_7 r(
xolos	SOlos	S o_7 l o s

This is the simplest tool because the input data can be any text in Spanish with punctuation marks and any number of words per line. It is assumed that the input text is at least, well written with all the orthographic rules of the Spanish language. This tool will predict where the tonic vowel of every word is with the help of our internal function called "vocal_tonica()". In the section of "Evaluation" we present a study of the accuracy of the vocal_tonica() function that works pretty well. Nevertheless one has to be careful with names, conjugated verbs and words that are not in Spanish like "ballet" or "cappuccino", because this function could fail.

One has also to put an eye on words with letter "x" because in Spanish, this letter has four different sounds as will be explained in the next section. The default sound that this tool will assume for all the words with "x" is /ks/ like in "sexto" or "examen" that sound like "seksto" and "eksamen" respectively but it will fail with words like "excepción" or "xavier" that sound like "esepción" and "javier" respectively.

B. Pronouncing Dictionary from Preprocessed Text

This tool provides more accurate PD's but it requires some preprocessing of the input data. When the input data comes from the transcription file of a corpus in Spanish, it is probably that you already have the text almost totally preprocessed as is needed by this tool. Anyway, you have to be sure that every line of the input file is between <s> </s> as in Fig. 2.

```
<s> mArca dOs trEs nuEve </s>
<s> llAma A trabAjo </s>
<s> mArca Ocho sEis trEs </s>
<s> llAma A escuEla </s>
```

Figure 2. Format required for the input file.

You also have to verify that the tonic vowel of every word is indicated by the same vowel but in upper case (see Fig. 2) and, as previously mentioned, you have to take into account the four different sounds or cases of the letter "x" by making the following substitutions:

- Letter "x" in words like "xochimilco", "xilófono" or "xochicalco" sounds like /s/ and it is substituted by "$" e.g. "$ochimIlco", "$ilOfono" and "$ochicAlco".
- Letter "x" in words like "xolos", "xicoténcatl" or "xoloescuincle" sounds like the Mexbet phoneme /S/ (see Table I). In those cases the "x" must be substituted by "S", e.g. "SOlos", "SicotEncatl" and "SoloescuIncle". This rule also applies for the conmbination of "s" and "h", like in "sharon" or "shanon". Those words must be transcribed as: "SAron" and "SAnon".
- Letter "x" in words like "mexico", "mexicali" or "xavier" sounds like letter "j" (phoneme /x/ in Mexbet, see Table I), so the "x" has to be substituted with a "J" like this: "mEJico", "meJicAli" and "JaviEr".
- Letter "x" in words like "exámen", "sexto" or "sexy" sounds like the phonemes /ks/ together, but

in this case; letter "x" remains unchanged, e.g. "exAmen", "sExto" and "sExy".

Table II shows some examples of words preprocessed correctly and then transcribed in Mexbet by this tool.

Notice that in Mexbet, the tonic vowels are marked with "_7".

C. Combine and Sort

As its name implies, this tool takes two plain UTF-8 text files as an input, and then, it combines them into one only file. The resulting file is sorted alphabetically.

This tool is perfect when one wants to incorporate missing entries, alternative pronunciations or elements of a filler dictionary to the final PD. A filler dictionary is like a regular PD with a word followed by its pronunciation. The difference is that the "word" is usually a non-speech event like noises, clicks, breath, laughter, etc.

One has to take into account that this tool does not eliminate duplicate entries in the resulting PD. In fact, this tool does not eliminate any entry at all and one has to be sure that there are no duplicate entries in the final file. This is especially important when one wants to incorporate alternative pronunciations to the PD. The alternative pronunciations are usually represented with the word, followed by a number in parenthesis as shown in Fig. 3.

```
SECRETARIO      s e g r( e t a r( i e
SECRETARIO(2)   s e k r( e t a r( i e
SECRETARIO(3)   s e k r( e t a r( i o
SECRETO s e k r( e t o
SECRETO(2)      t s e k k r( e t o
SECTOR  s e g t o r(
SECTOR(2)       s e k t o r
SECTOR(3)       s e k t o r(
SECTOR(4)       s e t o r(
SECTORES        e k t o r( e s
SECTORES(2)     s e k t o r( e s
SECTORIA        s e k t o r( i a
SECTORIAL       s e k t o r( i a l
SECTORIALES     s e g t o r( i a l e s
SECTORIALES(2)  s e k t o r( i a l e s
SECTORIALES(3)  s e t o r( i a l e s
SECUENCIA       s e k u e n s i a
SECUESTRADO     s e k u e s t r( a d o
```

Figure 3. PD with Alternative Pronunciations.

Notice that the word "SECTOR" is repeated many times but with different transcriptions, and the only way to differentiate between every entry containing that word is, the numbers in parentheses. So, one has to be careful while adding the alternative pronunciations to the PD. This process has to be made by hand because the online tools only produce canonical pronunciations.

III. Evaluation Experiments

The most important part of the tools presented in this document is the automatic phonetizer, incorporated in our internal function called T22(). An automatic phonetizer is a program which receives a word as an input, and returns its phonetic transcription. Our T22() function utilizes grapheme-to-phoneme rules in Spanish to calculate the phonetic transcriptions of the incoming words in Mexbet.

As previously mentioned, the tool that produces PD's from raw text, calculates the position of the tonic vowel of every word in the dictionary by using our vocal_tonica() function.

In this section, an evaluation of the vocal_tonica() and T22() functions will be presented.

A. Evaluation of the Vocal_Tonica() Function

For the evaluation of the vocal_tonica() function we utilized words extracted from the CIEMPIESS corpus. This database counts with 12155 tokens (or words with no repetitions). We took randomly 1539 of them that represents the 12.66% of the whole CIEMPIESS words. Then we eliminated the foreign words (87) and that is how we obtained a total of 1452 words to analyze.

We manually checked if they were correctly accentuated. The result is that 90.35% (1312 words) were correctly accented against 140 with a wrong position of their tonic vowels. Some of the reasons for these errors are that some words were conjugated verbs and names. Table III summarizes these results.

TABLE III. Evaluation of the Vocal_Tonica() Function

Words in the CIEMPIESS corpus	12155
Words Taken from the CIEMPIESS Corpus	1539
Number of Foreign Words Omitted	87
Number of Words Analyzed	1452
Wrong Accentuation	140
Correct Accentuation	1312
Percentage of correct Accentuation	**90.35%**

B. Evaluation of the T22() Function

For the evaluation of the T22() function we counted with two different comparison elements. The first one is the pronouncing dictionary of the DIMEx100 corpus[2] that counts with 11575 entries. The second one is the software called TRANSCRÍBEMEX that is mentioned with that name in [10], but it also appears in [12] and [13]. As a matter of fact, this pronouncing dictionary was made by human transcribers of the DIMEx100 corpus, aided by the TRANSCRÍBEMEX.

The TRANSCRÍBEMEX is a software tool with graphic interface coded in Perl that produces transcriptions in Mexbet, similar to the transcriptions produced by the T22() function but not identical.

In the present evaluation we compare the transcriptions of the TRANSCRÍBEMEX with the transcriptions generated by our T22() function.

A problem with the TRANSCRÍBEMEX is that it produces a set of symbols called "archiphonemes" ([10], [12], [13]): [-B], [-D], [-G], [-N], [-R]. An archiphoneme is a phonological symbol that groups several phonemes together. For example, [-D] is equivalent to any of the phonemes /d/ or /t/. To learn more about archiphonemes see [14].

Another problem was the words with the grapheme "x" that, as previously mentioned, they can have any of four

[2] Download the pronouncing dictionary of the DIMEx100 corpus at http://turing.iimas.unam.mx/~luis/DIME/

different pronunciations depending on the sound of the "x" in the current word. The TRANSCRÍBEMEX only manages the sound /ks/ for the grapheme "x". For that reason, we eliminated the words with "x" of the analysis. We also eliminated the alternative pronunciations.

Finally after all of these precautions, the result was that both tools are 99.2% similar, which means that our T22() function is reliable. Table IV summarizes the experiment and the results.

TABLE IV. COMPARISON BETWEEN TRANSCRÍBEMEX AND THE T22() FUNCTION.

Words in the DIMEx100 Corpus	11575
Alternative Pronunciations	2591
Words with the letter "x"	202
Archiphonemes	45
Number of Words Analyzed	8737
Non Identical Transcriptions	67
Identical Transcriptions	8670
Percentage of Identical Transcriptions	**99.2%**

IV. CONCLUTIONS

A set of novel, automatic, online and free access tools for producing pronouncing dictionaries for the Mexican Spanish has been presented.

The pronouncing dictionaries generated by these tools are useful in several fields of the speech processing.

The methodologies for using those tools properly have been explained and, an evaluation of the accuracy of them has demonstrated that they are reliable (above 90%).

ACKNOWLEDGMENT

The authors wish to thank UNAM PAPIIT/DGAPA project IT102314, CEP-UNAM and CONACYT for financial support, Nancy N. Martínez Gómez for her help on the test of the accuracy of the software tools, and Fréderick V. Álvarez Flores for his work on the upload and maintenance of the software tools in the CIEMPIESS-UNAM Project website.

REFERENCES

[1] D. Odden, *What is Phonology?* 1996.
[2] C. S. Salcedo, "The phonological system of Spanish," in Revista de Lingüística y Lenguas Aplicadas, Universitat Politècnica de València, 2010.
[3] T. N. Tomás, "El alfabeto fonético de la revista de filología Española," in *Anuario de Letras*, Northamptom, 1966, vol. 6, pp. 5–10.
[4] International Phonetic Association, *The Principles of the International Phonetic Association*, London: University College, 1949/1971.
[5] J. C. Wells, "SAMPA computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, 1997, chapter Part IV, Section B.
[6] J. Llisterri and J. B. Mariño, *Spanish Adaptation of SAMPA and Automatic Phonetic Transcription*, Technical Report of the ESPRIT PROJECT 6819, 1993.
[7] J. L. Hieronymus, *ASCII Phonetic Symbols for the World'S Languages: Worldbet*, Technical report. IPA, 1967. The principles of the International Phonetic Association.
[8] T. Lander and S. T. Metzler, *The CSLU Labeling Guide*, 1994.
[9] Tlatoa Group, *ASCII Phonetic Symbols for Mexican Spanish*, Universidad de las Américas, Mexico, 2000.
[10] J. Cuétara-Priede, "Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla," M.S. Thesis in spanish linguistics (in Spanish), Universidad Nacional Autónoma de México, Mexico, 2004.
[11] C. D. Hernández-Mena and J. A. Herrera-Camacho, "CIEMPIESS: A new open-sourced Mexican Spanish radio corpus," in *Proc. the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, 2014, pp. 371–375.
[12] L. A. Pineda, L. Villaseñor, J. Cuétara, H. Castellanos, and I. López, "DIMEx100: A new phonetic and speech corpus for Mexican Spanish," in *Lecture Notes in Computer Science*, Springer, 2004, pp. 974–984.
[13] L. A. Pineda, H. Castellanos, J. Cuétara, L. Galescu, J. Juarez, J. Llisterri, *et al.* "The corpus DIMEx100: Transcription and evaluation, " *Language Resources and Evaluation*, 2009.
[14] T. Akamatsu, "The theory of neutralization and the archiphoneme in functional phonology," in *John Benjamins Publishing*, 1988, vol. 43.

Carlos D. Hernández-Mena was born in Mexico City in 1983. He received his Bachelor's degree in the area of Electronics and Communication Engineering from The National Polytechnic Institute located in Mexico City in the year 2006. He received his M.E degree in the area of speech recogniton from the National Autonomous University of Mexico (UNAM) in 2010. His current phD research includes continuous speech recognition, microcomputers and applied phonetics. He assisted to the "Verano Científico Tec-Profesor" summer school in the Monterrey Institute of Technology and Higher Education (ITESM) located at Monterrey, Mexico in 2013. Nowadyas Carlos teaches microprocessors at UNAM. Prof Hernández-Mena is current member of the Acuctical Society of America (ASA), the Institute of Electrical and Electronics Engineers (IEEE) and the International Speech Communication Association (ISCA).

José A. Herrera-Camacho received degrees in Mechanical-Electrical Engineering, M. S. Electronic Engineering, and the Ph.D. Engineering, from Universidad Nacional Autónoma de México (UNAM), Mexico, in 1979, 1985, and 2001, respectively, the PhD was with support of the University of California in Davis. He did a postdoctoral research in 2001 at Carnegie Mellon University, and a sabbatical research at USC. He is author from more than 50 scientific papers on codification, recognition, and synthesis. He has contributed in a dozen projects between UNAM and companies about $50 million USD. He is coauthor of the book *Linear Algebra, theory and exercises*, printed 9 times from 1986 to 2009. Currently, he is the director of Speech Laboratory in the faculty of Engineering of UNAM.