A Set of Phonetic and Phonological Rules for Mexican Spanish Revisited, Updated, Enhanced and Implemented

Carlos-Daniel Hernández-Mena[a], Nancy-Norely Martínez-Gómez[b], and
José-Abel Herrera-Camacho[a]

[a]Departamento de Procesamiento Digital de Señales
Universidad Nacional Autónoma de México (UNAM)
ca_hernandez@uxmcc2.iimas.unam.mx
abelherrerac1@gmail.com

[b]Facultad de Filosofía y Letras, UNAM
nancy_luthien@hotmail.com

Abstract. This paper revisits the phonetic and phonological rules of grapheme-to-phoneme conversion for Mexican Spanish released in 2004 for a particular computational phonetic alphabet called "Mexbet", which was specially designed for the Spanish spoken in Central Mexico. Mexbet has proved its value over the years but it is still not well known. Implementation of these rules into real programming code will be explained, and the *fonetica2* library that is an open-source tool which allow users to develop customized phonetic transcription algorithms, will be presented.

Keywords: grapheme-to-phoneme convertion, mexican spanish, mexbet, phonetic alphabet.

1 Introduction

In the field of phonetics, it is necessary to register the sounds of human languages, which is the reason for the phonetic alphabet. A phonetic alphabet is a set of symbols that represents the sounds of speech. The process to convert a word written in a regular alphabet, to a particular phonetic alphabet is called "transcription". Some examples of phonetic alphabets are the RFE alphabet that is exclusive for the Spanish language [1] or the IPA alphabet that was created by the International Phonetic Asociation [2]. It contains the speech sounds of all the languages over the world.

The use of these alphabets have disadvantages in most fields of computer science because of something as simple as the character codification. The IPA, the RFE and other "classic" phonetic alphabets utilize symbols that are impossible to deal with in programming codes, usually written in plain text with an ASCII or UTF-8 set of characters. The solution to these problems is the creation of computational phonetic alphabets that are nothing but the translation

©L. Berthet, J. C. Chimal, E. A. Santos, E. Castillo
Advances in Computer Science
Research in computing Science 83, 2014 pp. 61-71

of classic alphabets into a set of symbols easy to incorporate in programming codes. Normally, these kind of alphabets are written with ASCII symbols and that is why we can call them by the full name of "ASCII Computational Phonetic Alphabets" or ACPA for short. Some examples of these ACPA could be the SAMPA alphabet that was designed for many languages including Spanish [3], the WORLDBET alphabet that is an ASCII adaptation of the IPA alphabet [4] or the OGIBET alphabet that was created by the Oregon Institute of Science and Technology and then adapted for the Mexican Spanish by the Tlatoa Group[1] [5].

Mexbet is another example of an ACPA that we will discuss in the present document. Mexbet is an ACPA specially designed for the Mexican Spanish, mostly based on WORLDBET but also incorporating OGIBET. It first appaeared in [6] as a part of a master thesis. After that, Mexbet went through transformations and updates over the years and has been published more than once [7], [8]. The current version of Mexbet in which we are interested, appeared in 2004 on another master thesis written by Javier Cuétara [9]. In this thesis, Cuétara introduced a number of changes and corrections to the current version of that time. All the data that he utilized to carry out his analysis came from the DIMEx100 corpus [10] which is a phonetic and speech corpus for Mexican Spanish created at IIMAS-UNAM[2].

Some examples where Mexbet was successfully used are the VOXMEX corpus [11] which utilized an early version of it, and the DIMEx100 corpus that used a newer and extended one, nevertheless these papers do not expose the transcripcion rules of Mexbet with much detail, actually, the most complete information is in the Cuétara thesis but a basic knowledge in phonetics is necessary to understand it well, and not to mention that the thesis is written in Spanish, which means that Mexbet rules will be less understandable for non-Spanish speaking researchers and engineers.

As we can see, there is a real need to have unified standards and free access language resources for a healthy development of speech technologies in Mexico and that is why the CIEMPIESS-UNAM Project[3] was created. The CIEMPIESS-UNAM Project aims to develop and share free and open-source tools for speech processing in the Spanish language. We have recently released our 17 hours database for speech recognition called CIEMPIESS [12] that uses Mexbet and it is completely free. We think that this may define new conditions for the growth and development of the field of speech processing in Mexico.

In the next sections we will sum up the full specification of Mexbet (as anyone has done before) that has demostrated its usefullness in the field of automatic speech recognition.

[1] Tlatoa is (or was) a Mexican speech technologies research group from the "Universidad de las Américas de Puebla" founded in 1997.

[2] IMASS-UNAM website at `http://www.iimas.unam.mx/`

[3] See `http://odin.fi-b.unam.mx/CIEMPIESS-UNAM/`

2 Phonological Rules of Mexbet

Phonology is the branch of linguistics which studies the speech sounds (phonemes) as an abstract system [13]. This means that in phonology the phonemes are represented only in a prototypic way. The branch of linguistics which tries to represent the speech sounds in the most accurate way as possible is phonetics [14].

In the 2004 version of Mexbet a total of 17 consonants and 5 vocals for the Mexican Spanish are found. This phonological level of Mexbet is called "T22" although it was decided to include two more phonemes that come from the Náhuatl language (also known as "Aztec" [15]) in this revision of Mexbet, because they are common in the Mexican Spanish. These phonemes are /S/ like in the word "xoloescuincle" (in English this phoneme could sound as "sh" like in "shannon") and /tl/ that is always at word endings like in "popocatépetl" or "citlaltépetl".

Table 1 shows the Mexbet symbols for the Mexican Spanish phonemes which provide information about the manner and the point of articulation of each one. This information is useful in speech recognition because it helps to group similar acoustic models together as the HTK toolkit requires [16].

Table 1. Phonolical Level of Mexbet also known as "T22"

	Points of Articulation						
Manners of Articulation	**Consonants**	**Labial**	**Labiodental**	**Dental**	**Alveolar**	**Palatal**	**Velar**
	Voiceless Stop	p		t			k
	Voiced Stop	b		d			g
	Voiceless Affricate					tS	
	Voiceless Fricative		f		s	S	x
	Voiced Fricative					Z	
	Nasal	m			n	n∼	
	Rhotic				r(/ r		
	Lateral				l	tl	
	Vowels				**Front**	**Central**	**Back**
	Close				i		u
	Mid				e		o
	Open					a	

In practice, a phonological transcription is nothing but a "transformation" of a word written in a conventional way (orthographic) into another that only includes the symbols of Table 1. Some of these transformations depends on the fact that many graphemes do not represent the speech sounds correctly, but they have to be written because of tradition or historical reasons, like the grapheme "h" that in spanish is usually mute, or the grapheme "z" that represents the phoneme /s/.

Table 2 shows the specific transformations of certain graphemes that were adopted in the *fonetica2* library that we will present in the following sections. In the case of the *fonetica2* library, it is assumed that every word is in Spanish, that is why transformations provided in Table 2 are sufficient and work pretty well. If you decide to do transcriptions of words in different languages like "cappuccino" or "ballet" it is probably that one will need to consider more transformations or the use of stop word lists. To see how to deal with grapheme transformations for different languages see [17].

Table 2. Transformations adopted to do phonological transcriptions in Mexbet

No ASCII Symbol : Example	**Transformation : Example**	**Ortographic Irregularity : Example**	**Phoneme Equivalence : Example**	**Ortographic Irregularity : Example**	**Phoneme Equivalence : Example**
“á” : “cuál”	“cuAl”	“cc” : “accionar”	/ks/ : “aksionar”	“gui” : “guitarra”	/g/ : “gitaRa”
“é” : “café”	“cafE”	“ll” : “llamar”	/Z/ : “Zamar”	“que” : “queso”	/k/ : “keso”
“í” : “maría”	“marIa”	“rr” : “carro”	“R” : “caRo”	“qui” : “quizá”	/k/ : “kisA”
“ó” : “noción”	“nociOn”	“ps” : “psicología”	/s/ : “sicologIa”	“ce” : “cemento”	/s/ : “semento”
“ú” : “algún”	“algUn”	“ge” : “gelatina”	/x/ : “xelatina”	“ci” : “cimiento”	/s/ : “simiento”
“ü” : “güero”	“gWero”	“gi” : “gitano”	/x/ : “xitano”	“y” (end of word) : “buey”	/i/ : “buei”
“ñ” : “niño”	“niNo”	“gue” : “guerra”	/g/ : “geRa”	“h” (no sound) : “hola”	“ola”

Substitutions in Table 2 are the first step to make phonological transcriptions in Mexbet. For example, non ASCII symbols like “ñ” or “ü” are substituted for ASCII symbols like “N” and “W” (respectively) to prevent character codification issues. Also the graphic accents on vowels (”á”, “é”, “í”, etc.) are substituted for the same vowel but in upper case. In general, at this point in the transcription process it is considered that vowels in upper case are tonic vowels even if they do not come from a vowel with a graphic accent. Other transformations of certain chains of graphemes (morphemes) that are shown in Table 2 are self-explanatory and the reason of choosing them are well discussed in [18].

There is a special problem with the grapheme “x” that in Mexican Spanish has 4 different sounds and there is no transcription rule for all of of them. Table 3 shows these four cases of the grapheme “x” and the substitution that has to be made in order to transcribe every case correctly.

Table 3. The four different sounds of the grapheme “x” in Mexican Spanish

Cases of “x”	**Phoneme Equivalence : Example**
“sexto”,“oxígeno”	/ks/ : “sexto”,“oksIgeno”
“xochimilco”,“xilófono”	/s/ : “sochimilco”,“silOfono”
“xolos”,“xicoténcatl”	/S/ : “Solos”,“SicotEncatl”
“ximena”,“xavier”	“j” : “jimena”,“javier”

If you want to make an automatic transcription of words containing the grapheme “x”, it is possible one has to consider the use of stop word lists. Notice that in words like “ximena” and “xavier” the substitution of the grapheme “x” is the grapheme “j” and not a phoneme like in the other cases. That is because in Mexbet, the phoneme /x/ is represented by the grapheme “j”. One has to be careful not to confuse the grapheme “x” with the phoneme /x/.

After applying transformations in Table 2 and Table 3, one may apply the last step of phonological transcription that is nothing but making the transformations of the missing graphemes that has not been touched until now. Table 4 shows these last transformations of graphemes into phomemes.

Table 4. Final grapheme to phoneme transformations

Grapheme	**Phoneme**	**Grapheme**	**Phoneme**
"A"	/a_7/	"ch"	/tS/
"E"	/e_7/	"c"	/k/
"I"	/i_7/	"j"	/x/
"O"	/o_7/	"v"	/b/
"U"	/u_7/	"z"	/s/
"N"	/n∼/	"r"	/r(/
"y" not at word ending	/Z/	"R"	/r/
"W"	/u/		

Notice that in Table 4 the Mexbet symbol "_7" indicates the tonic vowel. Graphemes like "p" or "t" are not in Table 4 because their phoneme equivalence are represented with the same symbols (/p/ , /t/ respectively).

The next section shows how to transform a phonological transcription in Mexbet T22 into a phonetic transcription in Mexbet level T50.

3 Phonetic Rules of Mexbet

Phonetics is the branch of linguistics that studies the speech sounds from the point of view of their acoustic realization and it is very interested in the position of the organs that produces the speech sounds (articulation features) [14]. 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.

Before applying the phonetic transcription rules it is necessary to have a good syllabification of the words that one wants to transcribe and for a good syllabification, it is usually needed to know where the tonic vowell is. Discussing the syllabification and the accentuation rules of the Spanish is out of the scope of the present article, but for more information about them, please see [19].

After having made a phonological transcription of a word, one can now transform it into a phonetic form. In the DIMEx100 corpus, it is evident that there is only one phonological level for Mexbet called T22, but there are two different levels of granularity in the phonetic version of Mexbet, the T44 and the T54[4]. Nevertheless, these levels does not contain all the Mexbet allophones that appeared in the 2004 version and they include symbols like "_c" for all the oclusive

[4] You can see all the versions of Mexbet used in the DIMEx100 corpus at `http://turing.iimas.unam.mx/~luis/DIME/CORPUS-DIMEX.html`

phonemes (e.g. /k_c/, /p_c/, /t_c/)[5]. In this revision, we want to introduce a new phonetic level of Mexbet called "T50" that contains all the phonemes and allophones of the version 2004 (except the phoneme /m_n/) plus the two Aztec phonemes /S/ and /tl/ and it does not contain symbols for closure moments.

The following list shows all the phonetic rules of Mexbet. They can allow users to perform accurate transcription algorithms. The Mexbet symbols in / / denote the phonemes and the symbols in [ ] denote the allophones. In this list we can find examples of some words transcribed in the T22 and the T50 level that illustrate well the corresponding rules. Transcriptions in / / are in T22 and transcriptions in [ ] are in T50. Words in " " are the words that we want to transcribe and we (manually) indicate the tonic vowel of each one with a capital letter (eg. "teclAdo", "ambulAncia", etc.).

1. **[a_j]: Followed by palatal consonant; In diphthong ending in /i/ e.g. "ayEr", "aIre" - / a . Z e_7 r(/ , / a_7 i . r(e / - [a_j . Z E_7 r(_\] , [a_j_7 i(. r(E] .**
2. **[a_2]: In diphthong ending in /u/; Followed by /o/; In closed syllable ending in /l/; Followed by /x/ e.g. "Aunque", "paOla", "altUra", "ajEno" - / a_7 u n . k e / , / p a . o_7 . l a / , / a l . t u_7 . r(a / , / a . x e_7 . n o / - [a_2_7 u(N . k_j e] , [p a_2 . o_7 . l a] , [a_2 l_[. t U_7 . r(a] , [a_2 . x e_7 . n o] .**
3. **/ a /: In every other context e.g. "cAsa" - / k a_7 . s a / - [k a_7 . s a] .**
4. / e /: In open syllable; In closed syllable ending in /m, n, s, d/ e.g. "pErro", "pensAr" - / p e_7 . r o / , / p e n . s a_7 r(/ - [p E_7 . r O] , [p e n . s a_7 r(_\] .
5. [E]: In closed syllable ending in a consonant (except /m, n, s, d/); in contact with /r/ or /r(/; In diphthong ending in /i/; Followed by /x/ e.g. "selvAtico", "mercAdo", "peinAr", "lejAno" - / s e l . b a_7 . t i . k o / , / m e r(. k a_7 . d o / , / p e i . n a_7 r(/ , / l e . x a_7 . n o / - [s E l . V a_7 . t i . k o] , [m E r(_\. k a_7 . D o] , [p E i(. n a_7 r(_\] , [l E . x a_7 . n o] .
6. **/ o /: In open syllable e.g. "comEr" - / k o . m e_7 r(/ - [k o . m E_7 r(_\] .**
7. **[O]: In closed syllable ending in any consonant; In contact with /r/ or /r(/; In diphthong ending in /i/; Followed by /x/ e.g. "montAr", "zOrra", "herOico", "mojAr" - / m o n . t a_7 r(/ , / s o_7 . r a / , / e . r(o_7 i . k o / , / m o . x a_7 r(/ - [m O n_[. t a_7 r(_\] , [s O_7 . r a] , [E . r(O_7 i(. k o] , [m O . x a_7 r(_\] .**
8. / i /: In open syllable e.g. "chiflAr" - / tS i . f l a_7 r(/ - [tS i . f l_0 a_7 r(_\] .
9. [I]: In closed syllable; In contact with /r/ or /r(/; Followed by /x/ e.g. "silbAr", "rIco", "quijAda" - / s i l . b a_7 r(/ , / r i_7 . k o / , / k i . x a_7 . d a / - [s I l . V a_7 r(_\] , [r I_7 . k o] , [k_j I . x a_7 . D a] .
10. [j]: At beginning of diphthong e.g. "diArio" - / d i a_7 . r(i o / - [d j a_7 . r(j o] .
11. [i( ]: At end of diphthong e.g. "sEis" - / s e_7 i s / - [s E_7 i(s] .
12. **/ u /: In open syllable e.g. "pulIr" - / p u . l i_7 r(/ - [p u . l I_7 r(_\] .**
13. **[U]: In closed syllable; In contact with /r/ or /r(/; Followed by /x/ e.g. "funcionAr", "puritAno", "sujEto" - / f u n . s i o . n a_7 r(/ , / p u . r(i . t a_7 . n o / , / s u . x e_7 . t o / - [f U n . s j o . n a_7 r(_\] , [p U . r(I . t a_7 . n o] , [s U . x e_7 . t o] .**
14. **[w]: At beginning of diphthong e.g. "cuEnto" - / k u e_7 n . t o / , [k w e_7 n_[. t o] .**
15. **[u( ]: At end of diphthong e.g. "reusAr" - / r e u . s a_7 r(/ - [r E u(. s a_7 r(_\] .**
16. / p /: In every context e.g. "repAso" - / r e . p a_7 . s o / - [r E . p a_7 . s o] .
17. **/ t /: In every context e.g. "tetEra" - / t e . t e_7 . r(a / - [t e . t E_7 . r(a] .**
18. [k_j]: Followed by front vowels e.g. "troquelAr" - / t r(o . k e . l a_7 r(/ - [t r(_0 O . k_j e . l a_7 r(_\] .
19. / k /: In every other context e.g. "cAza" - / k a_7 . s a / - [k a_7 . s a] .
20. **/ b /: At beginning of word; After nasal consonant e.g. "vAca", "sOmbra" - / b a_7 . k a / , / s o_7 m . b r(a / - [b a_7 . k a] , [s O_7 m . b r(a] .**
21. **[V]: In every other context e.g. "cabEza" - / k a . b e_7 . s a / - [k a . V e_7 . s a] .**
22. / d /: At beginning of word; After nasal consonant; After /l/ e.g. "dOna", "cOnde", "cElda" - / d o_7 . n a / , / k o_7 n . d e / , / s e_7 l . d a / - [d o_7 . n a] , [k O_7 n_[. d e] , [s E_7 l_[. d a] .
23. [D]: In every other context e.g. "ademAs" - / a . d e . m a_7 s / - [a . D e . m a_7 s] .

[5] This symbols represents the "closure moment" that is the time when the air is totally obstructed by the articulatory organs of the mouth when pronouncing an occlusive consonant.

24. **/ g /: At beginning of word; After nasal consonant e.g. “gAto”, “angOsto” - / g a_7 . t o / , / a n . g o_7 s . t o / - [g a_7 . t o] , [a N . g O_7 s_[. t o] .**
25. **[G]: In every other context e.g. “pegAr” - / p e . g a_7 r(/ - [p e . G a_7 r(_\] .**
26. / tS /: In every context e.g. “mochIla” - / m o . tS i_7 . l a / - [m o . tS i_7 . l a] .
27. **/ f /: In every context e.g. “refOrma” - / r e . f o_7 r(. m a / - [r E . f O_7 r(_\. m a] .**
28. [z]: Followed by voiced consonants e.g. “asbEsto” - / a s . b e_7 s . t o / - [a z . V e_7 s_[. t o] .
29. [s_[ ]: Followed by voiceless dental consonant e.g. “asterIsco” - / a s . t e . r(i_7 s . k o / - [a s_[. t E . r(I_7 s . k o] .
30. [z_[ ]: Followed by voiced dental consonant e.g. “dEsde” - / d e_7 s . d e / - [d e_7 z_[. D e] .
31. / s /: In every other context e.g. “recEta” - / r e . s e_7 . t a / - [r E . s e_7 . t a] .
32. **/ x /: In every context e.g. “mejOr” - / m e . x o_7 r(/ - [m E . x O_7 r(_\] .**
33. [dZ]: At beginning of word; After nasal consonant; After /l/ e.g. “yUnque”, “inyectAr”, “ulyAna” - / Z u_7 n . k e / , / i n . Z e k . t a_7 r(/ , / u l . Z a_7 . n a / - [dZ U_7 N . k_j e] , [I n_j . dZ E k . t a_7 r(_\] , [U l_j . dZ a_7 . n a] .
34. / Z /: In every other context e.g. “payAso” - / p a . Z a_7 . s o / - [p a_j . Z a_7 . s o] .
35. **/ m /: In every context e.g. “medievAl” - / m e . d i e . b a_7 l / , [m e . D j e . V a_2_7 l] .**
36. **[m_n]: After /n/. Not implemented because it modifies the syllabication e.g. “enmEdio” - “ e n . m é . d i o “ - / e n . m e_7 . d i o / - [e . m_n e_7 . D j o] .**
37. [m]: Followed by labial consonant e.g. “inviErno” - / i n . b i e_7 r(. n o / - [I m . b j E_7 r(_\. n o] .
38. [M]: Followed by labiodental consonant e.g. “infOrme” - / i n . f o_7 r(. m e / - [I M . f O_7 r(_\. m e] .
39. [n_[ ]: Followed by dental consonant e.g. “anteriOr” - / a n . t e . r(i o_7 r(/ - [a n_[. t E . r(j O_7 r(_\] .
40. / n /: At beginning of word or syllable e.g. “tIna” - / t i_7 . n a / - [t i_7 . n a] .
41. [n_j]: Followed by palatal consonant e.g. “inyecciOn” - / i n . Z e k . s i o_7 n / - [I n_j . dZ E k . s j O_7 n] .
42. [N]: Followed by velar consonant e.g. “hangAr” - / a n . g a_7 r(/ - [a N . g a_7 r(_\] .
43. **/ n~ /: In every context e.g. “nIño” - / n i_7 . n~ o / - [n i_7 . n~ o] .**
44. [l_[ ]: Followed by dental consonant e.g. “Alto” - / a_7 l . t o / - [a_2_7 l_[. t o] .
45. [l_j]: Followed by palatal consonant e.g. “salchIcha” - / s a l . tS i_7 . tS a / - [s a_2 l_j . tS i_7 . tS a] .
46. [l_0]: After /p, k, f/ e.g. “plAnta” - / p l a_7 n . t a / - [p l_0 a_7 n_[. t a] .
47. / l /: In every other context e.g. “lOco” - / l o_7 . k o / - [l o_7 . k o] .
48. **[r(_0]: After /p, t, k, f/ e.g. “crEma” - / k r(e_7 . m a / - [k r(_0 E_7 . m a] .**
49. **[r(_\]: At the end of word or syllable e.g. “mermAr” - / m e r(. m a_7 r(/ - [m E r(_\. m a_7 r(_\] .**
50. **/ r(/: In every other context e.g. “arEna” - / a . r(e_7 . n a / - [a . r(E_7 . n a] .**
51. / r /: At beginning of word or syllable; After /s, n, l/ e.g. “honrAdo” - / o n . r a_7 . d o / - [O n . r a_7 . D o] .
52. **/ S /: In every context e.g. “xicotEncatl” - / S i . k o . t e_7 n . k a . tl / - [S i . k o . t e_7 N . k a_j . tl] .**
53. / tl /: At end of word e.g. “popocatEpetl” - / p o . p o . k a . t e_7 . p e . tl / - [p o . p o . k a . t e_7 . p e . tl] .

4 Online Tools and Evaluation

As previously mentioned, the *fonetica2* library is an open-source software tool that contains functions to transcribe Spanish words into a phonetic and a phonological level[6]. The *fonetica2* library is coded in Python and includes a total of six functions that can be easily incorporated to the user code who can see them as black boxes. Each of these functions accept Spanish words in lowercase as arguments. These words can have the tonic vowel marked in uppercase or not (e.g. cAldo , aviOn , comida , etc.).The functions are:

– **vocal_tonica()**: Returns the same incoming word but with its tonic vowel in uppercase (e.g. cAsa, pErro, gAto, etc.).

[6] Download the *fonetica2* library at `http://www.ciempiess.org/downloads`

- **div_sil()**: Returns the syllabification of the incoming word.
- **TT()**: "TT" is the acronym for "Text Transformation". This function produces the text transformations in Table 2 over the incoming word. All of them are perfectly reversible.
- **TT_INV()**: Produces the reverse transformations made by the TT() function.
- **T22()**: Produces a phonological transcription in Mexbet T22 of the incoming word.
- **T50()**: Produces a phonetic transcription in Mexbet T50 of the incoming word.

In the website of the CIEMPIESS-UNAM Project you can also find four online tools to test the functions: vocal_tonica(), div_sil(), T22() and T50(). You only need to write a Spanish word in a textbox, press a button and see the result provided.

4.1 Evaluation of the vocal_tonica() function

For the evaluation of the vocal_tonica() function we used words extracted from the CIEMPIESS corpus. This database counts with 12155 tokens (or words with no repetitions). We took randomly 1539 words that is 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.

4.2 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 that counts with 11575 entries. The second one is the software called TRANSCRÍBEMEX that is mentioned with that name in the Cuétara thesis. 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 phonological transcriptions in Mexbet T22, phonetic transcriptions in Mexbet T54 and shows the syllabification of each word. It can be used to transcribe entire sentences and not just isolated words.

We decide to use only the transcriptions in T22 of the TRANSCRÍBEMEX to compare with our T22() function because the transcriptions in T54 are so different and even incomplete with respect to the transcriptions produced by our T50() function.

A problem with the TRANSCRÍBEMEX is that it produces a set of symbols called "archiphonemes" [20]: [-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/.

Another problem were the words with the grapheme "x" that, as previously mentioned, can have any of four different pronunciations depending on the sound of the "x" in the current word. The TRANSCRÍBEMEX only utilizes the sound /ks/ for the grapheme "x". For that reason, words with "x" in the analysis were eliminated.

The other element that was eliminated were the alternative pronunciations that in the DIMEx100 corpus appear with a digit in parentheses (e.g. DAÑADOS d a n~ a d o s ; DAÑADOS(2) d a n~ a o s). These alternate pronunciations were created by human transcribers based on the recordings that they had to transcribe.

Finally after all of these precautions, we compared the transcriptions of the TRANSCRÍBEMEX with the transcriptions of the T22() function. The result was that both tools are **99.2%** similar which means that our T22() function is reliable.

4.3 Indirect Evaluations

For the evaluation of the TT(), and the div_sil() functions we did not have elements of comparison available as with the T22() function, nevertheless, we did not need them.

As we have demonstrated, the T22() function is reliable with respect to the TRANSCRÍBEMEX that is also reliable. Then we assume that the TT() and the div_sil() functions works well because the T22() needs them to work, and if they were wrong, the transcriptions of the T22() would be wrong too.

We did not have comparison elements for the evaluation of the T50() function either, so we manually had to test hundreds of words in alpha versions of the *fonetica2* library. The result is that all the 74 example words in the phonetic rules listed in the previous section were generated with no errors using our T50() function.

5 Conclutions

We have presented a set of phonological and phonetic rules for Mexican Spanish in English, we have implemented them, we have evaluated them and we have demonstrated that they are reliable and we hope that this contribution improves the access to them for researchers and engineers all over the world.

We also have shown novel and open-source tools created by the CIEMPIESS-UNAM Project that can contribute for the development of speech technologies in Mexico and other countries.

Acknowledgements

We thank UNAM PAPIIT/DGAPA project IT102314, CEP-UNAM and CONACYT for finantial support. We thank Professor Javier Cuétara for his advice on the adaptation of these rules into real programming code, Ivan Vladimir Meza Ruiz for providing us the pronouncing dictionary of the DIMEx100 corpus, Alejandra Chavarría for her help on the test of an alpha version of the *fonetica2* library and Tara Hansen for her revision to the present manuscript.

References

1. Navarro-Tomás, T.: El alfabeto Fonético de la Revista de Filología Española. In: Anuario de Letras, vol. 6, pp. 5–10, Northamptom, Massachusetts (1966)
2. International Phonetic Association.: The principles of the International Phonetic Association. London, University College (1949/1971)
3. Llisterri, J., Mariño, J.B.: Spanish Adaptation of SAMPA and Automatic Phonetic Transcription. In: Espirit Project Technical Report, vol. 6819 (1993)
4. Hieronymus, J.L.: ASCII Phonetic Symbols for the World's Languages: Worldbet. J. of the Phonetic Association. 23 (1993)
5. Kirschning, I.: Research and Development of Speech Technology & Applications for Mexican Spanish at the Tlatoa Group. In: CHI'01 Extended Abstracts on Human Factors in Computing Systems, pp. 49–50, ACM (2001)
6. Uraga, E.: Modelado Fonético para un Sistema de Reconocimiento de Voz Continua en Español. Master's thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey-Campus Morelos, Mexico (1999)
7. Uraga, E., Pineda, L.A.: A Set of Phonological Rules for Mexican Spanish. IIMAS, UNAM (2000)
8. Uraga, E., Pineda, L. Automatic Generation of Pronunciation Lexicons for Spanish. In: CICLING, pp. 330-338. Springer, Heidelberg, Berlin (2002)
9. Cuétara-Priede, J: Fonética de la ciudad de México Aportaciones desde las Tecnologías del Habla. Master's thesis in Spanish Linguistics, UNAM (2004)
10. Pineda, L. A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Villaseñor, L.: The Corpus DIMEx100: transcription and evaluation. In: LREC, vol. 44, no. 4, pp. 347–370 (2010)
11. Uraga, E., Gamboa, C.: VOXMEX Speech Database: Design of a Phonetically Balanced Corpus. In: LREC (2004)
12. Hernández-Mena, C.D., Herrera-Camacho, J.A.: CIEMPIESS: A New Open-Sourced Mexican Spanish Radio Corpus. In: Calzolari, N., Choukri, K. , et al (eds.), LREC'14. pp. 371-375. ELRA, Reykjavik, Iceland (2014)
13. Salcedo, C.S.: The Phonological System of Spanish. In: Revista de Lingüística y Lenguas Aplicadas. Universitat Politecnica de Valencia (2010)
14. Odden, D.: What is Phonology? (1996)
15. Clavigero, F.S.: Rules of the Aztec Language: Classical Nahuatl Grammar. University of Utah Press (1973)
16. Hernndez-Mena, C.D., Herrera-Camacho, J.A.: Creating a Grammar-Based Speech Recognition Parser for Mexican Spanish Using HTK, Compatible with CMU Sphinx-III System. I. J. of Electronics and Electrical Engineering. vol. 3, no. 3, pp. 220–224. February (2015)
17. Daelemans, W. M., Van den Bosch, A. P.: Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion. In: Progress in Speech Synthesis, pp. 77–89. Springer, New York (1997)
18. Ríos-Mestre, A.: La Información Lingüística en la Transcripción Fonética Automática del Español. In: Procesamiento del Lenguaje Natural, no. 13, pp. 381–387 (1993)
19. Quilis, A.: Tratado de Fonología y Fonética Españolas. Editorial Gredos (1993)
20. Pineda, L.A., Pineda, L.V., Cuétara, J., Castellanos, H., López, I.: DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. In: IBERAMIA, pp. 974–983. Springer, Berlin, Heidelberg (2004)