# CIEMPIESS: A New Open-Sourced Mexican Spanish Radio Corpus

**Carlos D. Hernández-Mena, Abel Herrera-Camacho**

Departamento de Procesamiento Digital de Señales
Universidad Nacional Autónoma de México (UNAM)
ca_hernandez@uxmcc2.iimas.unam.mx, abelhc@hotmail.com

**Abstract**

This paper presents the development of the "Corpus de Investigación en Español de México del Posgrado de Ingeniería Eléctrica y Servicio Social" (CIEMPIESS) that is a new open-sourced corpus extracted from Spanish spoken FM podcasts in the dialect of the center of Mexico. The CIEMPIESS corpus was designed to be used in the field of automatic speech recongnition (ASR) and it is provided with two different kind of pronouncing dictionaries, one of them containing the phonemes of Mexican Spanish and the other containing this same phonemes plus allophones. Corpus annotation took into account the tonic vowel of every word and the four different sounds that letter "x" presents in the Spanish language. CIEMPIESS corpus is also provided with two different language models extracted from electronic newsletters, one of them takes into account the tonic vowels but not the other one. Both the dictionaries and the language models allow users to experiment different scenarios for the recognition task in order to adequate the corpus to their needs.

**Keywords:** spanish radio corpus, mexican spanish corpus, mexican phonemes, mexican allophones

## 1. Motivation

Nowadays Mexican Spanish remains a resource-scarce language, but this lack of resources is not exclusive for this particular dialect, in general, development of tools for ASR in languages other than English is not always easy and depends on the the language you want to recognize. For example, You can use the CMU-SLMTK[1] to create a language model for your application but not for creating the pronouncing dictionary, because it is only generated with the English phonemes and there is no other similar widespread use tool for every language you may want.

Hence, when you want to recognize other languages, you also have to select the appropriate set of phonemes. In case of Spanish language, you can choose between different computational phonetic alphabets like SAMPA (Wells, 1997) or Worldbet (Hieronymus, 1994), but you have to be careful because these alphabets are created for several languages and dialects, and you may have troubles if you do not have basic knowledge of phonetics. Ideally, for many engineers and computational scientists it would be better if they could have only the set of phoneme and allophones they need with no worries of phonetic issues. Give solutions for that is usually responsibility of researchers and specialists into their own countries.

In the literature you can find some few corpus for the Spanish language (see (Llisterri, 2004)), but you have to adapt them to the dialect of Mexico (you can see an example of this kind of adaptation in (Varela et al., 2003)) if you want the best results.

This kind of adaptation issues and the scarce resources for the variant of the Spanish spoken in the center of Mexico is our main motivation for the development of the CIEMPIESS corpus that is an open-source tool designed for the creation of acoustic models for ASR systems.

We argue that creation of CIEMPIESS corpus as an open-source tool is not altruism, but it is a real need for development of speech technologies for the particular needs of our country. We can find other examples of these kind of "altruistic" ideas on the creation of the operating system "Linux"."Linux" was created for the necessity of having an open-source operating system for research and at this time it is supported for thousands of programmers all over the world and it is totally free! (you can read two interesting articles of what motivates people to develope free software in (Hars and Ou, 2001; Hertel et al., 2003)).

## 2. Corpus

CIEMPIESS is a radio corpus in the Mexican Spanish spoken at the center of Mexico, specifically at Mexico City. This is an important detail because "Mexico's City population is representative of the whole country" as we can see in (Pineda et al., 2009).

The total extension of CIEMPIESS is 17 hours.

CIEMPIESS has been annotated at the word level and tonic vowels have been considered in the transcription files. It is provided the language models and pronouncing dictionaries in order to increase its flexibility for the recognition task as we will see in the following sections.

### 2.1. Utterances

CIEMPIESS corpus has been taken from 43 one-hour duration FM radio programs[2], recorded in MP3 stereo format, using a 44.1 kHz sample rate and a bit-rate of 128 kbps or higher.

From these recordings were selected just the utterances considered "clean" that means that the utterances should be made by one only person with no background noises, whispers, music, foreign accents, white noise or static. We based our idea of "clean speech" partially on (Ostendorf et al., 1995). After that, the utterances were transformed into 16,717 16-bit audio files using a sampling rate of 16 kHz in the NIST Sphere PCM mono format.

[1] Statistical Language Modeling Toolkit by Carnegie Mellon University. See http://www.speech.cs.cmu.edu/SLM/toolkit.html

[2] Downloaded from http://podcast.unam.mx/

Every file name contains information about the gender of every speaker.

The total contribution of male voices is 77.86% against 22.14% of female voices. This means that the corpus is not gender balanced. Other examples of gender unbalanced corpus are found in (Wang et al., 2005; Federico et al., 2000)

### 2.2. Test Set

The 16,717 utterances which conforms the entire corpus sum a total of exactly 17.5 hours, so we subtracted 700 utterances to create the test set and the remaining 16017 utterances sum the 17 hours mencioned above.

Aditionally to these 700 utterances we include other 300 extracted from interviews, broadcast news and read speech. It is important to specify that the utterances selected for the test set are not considered in the training task of any of the recognition experiments presented in this paper. Test set is not considered at the language model either.

### 2.3. Language Model

The language model provided with CIEMPIESS was taken from "Boletines UNAM" [3] which is an electronic newsletter published by UNAM. All of the articles collected to create the language model were published between January of 2010 to February of 2013 and its total size is about 1.5 millions of words. Any of the utterances of CIEMPIESS were included in the language model. We also provide two different language models with the corpus. One of them includes the total words in lowercase and the other includes the tonic vowel of every word in uppercase. We used the CMU-SLMTK (Rosenfeld, 1995) to create the language model in the ARPA[4] format and the "DUMP file" required for doing experiments with the recognition system CMU-SPHINX (CMU, 2006).

### 2.4. Pronouncing Dictionaries

A previous step on the creation of the pronouncing dictionaries was the indication of tonic vowels. To do so, we create an automatic tool, base on the accentuation rules for the spanish language listed in (Quilis, 1999). To create the dictionary we programed an automatic phonetizer[5] based on grapheme-to-phoneme rules in (Cuetara-Priede, 2004).

We provide exactly four different pronouncing dictionaries with the corpus. One of them with the phonemes of the Mexican Spanish, other with this same phonemes plus a set of the most common allophones (see (Cuetara-Priede, 2004)) and both of them in a "no tonic" version. Every dictionary has about 50000 words.

### 2.5. Segmentation

CIEMPIESS corpus was annotated at the word level using PRAAT (Boersma and Weenink, 2013) which produces a kind of text files (segmentation files) containing the information of the beginning and ending of every word into every utterance. These segmentation files has the extension "textgrid".

We provide segmentation files for the 16717 utterances of the CIEMPIESS corpus but not for the remaining 300 utterances of the test set.

## 3. Phonetic Alphabet

The computational phonetic alphabet chosen for the pronouncing dictionaries was "Mexbet", that is based on Worldbet (Hieronymus, 1994) and it was specially designed for the Mexican Spanish. Mexbet was first presented in (Cuetara-Priede, 2004) and it was used at corpus DIMEx100 that is a Mexican Spanish corpus created in 2005 at UNAM (see (Pineda et al., 2004; Pineda et al., 2009)).

Mexbet counts with different levels of granularity. The basic one is called "T22" that includes the 22 phonemes for the Mexican Spanish. Aditionally to this 22 phonemes, we include the phoneme with the IPA[6] symbol /ʃ/. This phoneme is taken for the "náhuatl", that is an indigenous language to the region of Mesoamerica (Mexico included). We decided to include the phoneme /ʃ/ not because of its ocurrence (that is quite low) but because it is one of the four sounds that letter "x" presents in the Spanish language. Henceforth, when we mention T22 we will refer to the 22 phonemes plus the phoneme /ʃ/.

Table 1 shows the symbols for Mexbet T22 and their equivalents in IPA.

| 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 |
| ɟ | Z | u | u |
| m | m | | |

Table 1: Symbols for Mexbet T22.

Additionally to this T22, Mexbet includes other levels of granularity [7] and they were used at the DIMEx100 Corpus. Those are T54 and T44, but none of these levels include all the allophones contained at the full Mexbet. For that reason we chose our own level of granularity and we named it "T50".

Table 2 shows the symbols for Mexbet T50 and their equivalents in IPA.

[3] See http://www.dgcs.unam.mx/boletin/bdboletin/basebol.html

[4] For a good explanation of the APRA fotmat see http://www1.icsi.berkeley.edu/Speech/faq/grammarfmts.html

[5] Both the phonetizer and the accentuation tool were coded in Python.

[6] International Phonetic Alphabet

[7] See http://turing.iimas.unam.mx/∼luis/DIME/CORPUS-DIMEX.html

| IPA | Mexbet | IPA | Mexbet | IPA | Mexbet |
|---|---|---|---|---|---|
| p | p | s̬̪ | z_[ | ɾ | r( |
| t | t | s̬ | z | r | r |
| $k^j$ | k_j | ʝ | Z | ɾ̥ | r(_0 |
| k | k | ɣ̞ | G | ɹ | r(_\ |
| b | b | m | m | j | j |
| d | d | $\overset{n}{m}$ | m_n | i̯ | i( |
| g | g | ɱ | M | i | i |
| t͡ʃ | tS | n̪ | n_[ | i̞ | I |
| d͡ʒ | dZ | n | n | e | e |
| f | f | $n^j$ | n_j | e̞ | E |
| s̪ | s_[ | ɲ | n∼ | $a^+$ | a_j |
| s | s | $n^ɣ$ | N | a | a |
| x | x | l̪ | l_[ | a̱ | a_2 |
| ʃ | S | l | l | o | o |
| β̞ | V | $l^j$ | l_j | o̞ | O |
| ð̞ | D | l̥ | l_0 | u̯ | u( |
| u | u | u̞ | U | w | w |

Table 2: Symbols for Mexbet T50.

There is missing the symbol "_7" in both tables that is used to indicate the tonic vowel. In the next section we will show some examples of how to indicate tonic vowels using this symbol.

## 4. Annotation Methodology

As we have mentioned above, CIEMPIESS corpus was annotated at the word level. Firstly, all the utterances were transcribed orthographically using several conventions that we will explain at this section. The reason for this, is that all the words in the transcription file need to be transformed ("pre-transcription step" ) in order to obtain a correct phonetization of them.

After pre-transcription, we automatically generated the segmentation files for PRAAT. Then, those segmentation files were aligned to the utterances, and the annotation process is then finished.

### 4.1. Pre-Transcription

This "pre-transcription" process is a necessary step for the automatic phonetizer, because in the Spanish language you need to know where the tonic vowel is to do a correct syllabication and you need a correct syllabication to do a correct phonetic transcription. We indicate the tonic vowel with a capital letter (ej. "pErro", "gAto", etc.).

Table 3 shows examples of words with their pre-transcriptions and their phonetic transcriptions in T22.

### 4.2. Substitution of "ñ" and "x"

Other convention in the pre-transcription process is to convert every "ñ" into "N" in order to avoid codification problems.

After that, the correct indication of the the different sounds of the letter "x" is a very important step in the pre-transcription process as we explain below.

In Mexican Spanish the letter "x" has four different sounds but we do not have an automatic tool to indicate them, so we have to do it by hand, following these simple heuristic rules:

| Word | Pre-Transcription | Mexbet T22 |
|---|---|---|
| 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 | peNAsco | 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 |
| gina | gIna | Z i_7 n a |
| correo | corrEo | k o r e_7 o |
| sharon | SAron | S a_7 r( o n |
| sexenio | seKSEnio | 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 |

Table 3: Examples of transcriptions in Mexbet T22.

1. Letter "x" in words like "exámen", "sexto", "sexy", sounds like /ks/ and it is substituted by "KS", so the correct pre-transcription of those words is: "eKSAmen", "sEKSto" and "sEKSy".

2. Letter "x" in words like "xochimilco", "xilófono", "xochicalco" sounds like /s/ and it is substituted by "$". Ej. "$ochimIlco", "$ilOfono" and "$ochicAlco".

3. Letter "x" in words like "xolos", "xicoténcatl" or "xoloescuincle" sounds like the phoneme /ʃ/ that corresponds with /S/ in MEXBET Alphabet (see table 1). In those cases the "x" must be substituted by "S": "SOlos", "SicontEncatl" and "SoloescuIncle". This rule also applies for the combination of "s" and "h", that is "sh", like in "sharon" or "shanon". Those words must be transcribed as: "SAron" and "SAnon".

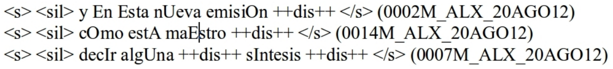
4. Letter "x" in words like "méxico", "mexicali" or "xavier" sounds like letter "j" (phoneme /x/ in IPA), so the "x" has to be substituted with a "J" like this: "mEJico", "meJicAli" and "JaviEr".

### 4.3. Transcription File

After pre-transcription and annotation processes are finished, we are able to create the the pronouncing dictionary (as we have explained in previous sections), and the trancription file.

The transcription file is in the format of the CMU-SPHINX speech recogntion system (CMU, 2006) (you can read more about this file format in section 4.9, step 8 of (Chan et al., 2007)). In this file format, you have to put the utterance between the labels <s></s>and you have to write the key of the audio file in parentheses as you can see in figure 1.

Note that in the examples of figure 1 appear symbols like ++dis++ and <sil>. The former is used to indicate disfluencies and the latter is used to indicate silences. What we understand by "disfluency" is non well-formed words, aspirations, dubitations (like "eeemmm, aaaamm, etc.), mispronunciations, or even noises that appeared for a short time in



```
<s> <sil> y En Esta nUeva emisiOn ++dis++ </s> (0002M_ALX_20AGO12)
<s> <sil> cOmo estA maEstro ++dis++ </s> (0014M_ALX_20AGO12)
<s> <sil> decIr algUna ++dis++ sIntesis ++dis++ </s> (0007M_ALX_20AGO12)
```

Figure 1: Example of some lines in the transcription file.

the record but does not affect the entire utterance. The silences are easier to identify because they appear between words and dont contain any noise or speech.

## 5. Experiments

In order to do some verification of all the files provided with the corpus, we performed four recognition experiments using SPHINX 3. We aimed to determine wich conditions are best for the recognition task plotting a learning curve for each one. These "conditions"refer to the use of the different pronouncing dictionaries and language models included with the corpus. All the experiments were based on three-state HMM models. In our particular case, rule of thumb suggests that 1000 recordings are equivalent to 1 hour. According to that, we stimate the number of senones on Table 4 that is based on the SPHINX-3 FAQ[8].

| No. of Recordings | No. of senones |
|---|---|
| 1000-3000 | 500 |
| 4000-6000 | 1000 |
| 6000-8000 | 2500 |
| 8000-10000 | 4000 |
| 10000-17000 | 5000 |

Table 4: Numbers of senones/Number of recordings.

Figure 2 shows four different learning curves with the following conditions:

1. **T22 TONICS**: Even the language model and the pronouncing dictionary consider the tonic vowel of every word and the granularity of phonetization is T22.
2. **T22 NO TONICS**: language model and pronouncing dictionary does not consider tonic vowels and the granularity of phonetization is T22.
3. **T50 TONICS**: Even the language model and the pronouncing dictionary consider the tonic vowel of every word and the granularity of phonetization is T50.
4. **T50 NO TONICS**: language model and pronouncing dictionary does not consider tonic vowels and the granularity of phonetization is T50

As we can see in figure 2, the best training conditions for the recognition task were "T22 TONICS", however, the difference between "T22 TONICS" and "T22 NO TONICS" was too small (just 1.7%). "T22 TONICS" was also the more stable curve. The final word error rate (WER) scores in figure 2 are:

- T22 TONICS = 44.0% (WER)
- T22 NO TONICS = 45.7% (WER)
- T50 TONICS = 50.5% (WER)
- T50 NO TONICS = 48.0% (WER)

We can also see that all the learning curves are consistent with the idea of "the more you train, the more you recognize" and the curves never converge. That means that we need to add more utterances for better performance.

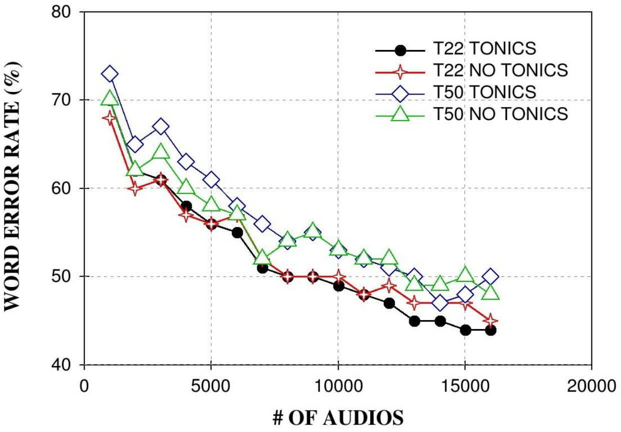


Figure 2: Learning Curves for different training conditions.

## 6. Conclusion

We conclude that in many senses, the Spanish spoken in Mexico is still a scarce resources language and it is responsibility of Mexican researchers and specialists to modify that reality. We assumed that responsibility with the creation of CIEMPIESS that it is unique in its type (size and availability).

Based on results of the experiments we also concluded that efforts made by marking the tonic vowel of every word could be good to improve the system performance, but we still need to do more experiments to prove it, However, we consider that the CIEMPIESS corpus is a useful tool for developing recognition systems and it represents a seed, and it could result in better and better things.

CIEMPIESS is ready to be used by the community, it certainly has the size enough to do many types of experiments not just for the recognition task only, but for linguistics, speaker identification, expert systems, and so on.

## 7. Further Work

We will do efforts for becoming CIEMPIESS into a gender balanced corpus and we will grow it size to 100 hours at least. We will also work with the language models. We need to have better control of the entrophy and we can create different kinds of language models for different kinds of speech. We will study the role that tonic vowels play in the recognition task and we will investigate better and faster ways to create corpus like the CIEMPIESS.

[8] See http://www.speech.cs.cmu.edu/sphinxman/FAQ.html

## 8. Acknowledgements

We thank UNAM PAPIIT/DGAPA project IT102314, CEP-UNAM and CONACYT for finantial support.

Special thanks to Jorge Valeriano Assem and Elena Vera for the support they give to the social service program dedicated to the CIEMPIESS Project. We thank to Professor Javier Cuétara for his advices on the creation of the automatic phonetizer and we thank to Alejandra Chavarria for her assistance in detecting bugs on it. We also thank to Felipe Orduña for his criticism on the first version of this document and we thank to Ivan V. Meza for his observations and advices on the final version. And finally, we thank to all of the social service students who helped to the creation of the corpus.

## 9. References

Paul Boersma and David Weenink. 2013. Praat: doing phonetics by computer. Version 5.3.51 retrieved 2 June 2013 from http://www.praat.org/.

Arthur Chan, Evandro Gouvea, and Rita Singh, 2007. *The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources* , March.

CMU. 2006. The CMU Sphinx Open Source Speech Recognition Engines. Version 3.

Javier Cuetara-Priede. 2004. Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla . Msc. thesis in spanish linguistics. in Spanish.

Marcello Federico, Dimitri Giordani, and Paolo Coletti. 2000. Development and Evaluation of an Italian Broadcast News Corpus. In *LREC*. European Language Resources Association.

Alexander Hars and Shaosong Ou. 2001. Working for free? - Motivations of participating in Open Source Projects. In *The 34th Hawaii International Conference on System Sciences*.

G. Hertel, S. Niedner, and S. Herrmann. 2003. Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research Policy*, 32(7):1159–1177.

J.L. Hieronymus. 1994. ASCII phonetic symbols for the world's languages: Worldbet . Technical report.

Joaquim Llisterri. 2004. Las tecnologías del habla para el español . pages 123–141. Fundación Española para la Ciencia y la Tecnología.

M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. Technical report ecs-95-001.

Luis Alberto Pineda, Luis Villaseñor Pineda, Javier Cuétara, Hayde Castellanos, and Ivonne López. 2004. DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. In Christian Lemaître, Carlos A. Reyes García, and Jesús A. González, editors, *IBERAMIA*, volume 3315 of *Lecture Notes in Computer Science*, pages 974–984. Springer.

Luis A. Pineda, H. Castellanos, Javier Cuetara Priede, L. Galescu, J. Juarez, Joaquim Llisterri, P. Prez-Pavn, and Luis Villaseñor. 2009. The Corpus DIMEx100: Transcription and Evaluation . Language Resources and Evaluation.

Antonio Quilis. 1999. *Tratado de Fonología y Fonética Españolas*. Ed. Gredos, second edition edition.

R. Rosenfeld. 1995. The CMU statistical language modelling toolkit and its use in the 1994 ARPA CSR evaluation. In *Proceedings of the ARPA Spoken Language Technology Workshop*, Austin TX.

Armando Varela, Heriberto Cuayáhuitl, and Juan Arturo Nolazco-Flores. 2003. Creating a Mexican Spanish Version of the CMU Sphinx-III Speech Recognition System. In Alberto Sanfeliu and José Ruiz-Shulcloper, editors, *CIARP*, volume 2905 of *Lecture Notes in Computer Science*, pages 251–258. Springer.

H. M. Wang, B. Chen, J. W. Kuo, and S. S. Cheng. 2005. MATBN: A Mandarin Chinese broadcast news corpus. *Int. Journal of Computational Linguistics and Chinese Language Processing*, 10(2):219–236.

J. C. Wells. 1997. SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*, chapter Part IV, Section B.