Creating Mexican Spanish Language Resources through the *Social Service* Program

Carlos Daniel Hernández-Mena, Ivan Vladimir Meza Ruiz
Language and Voice Lab, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Reykjavik University, Universidad Nacional Autónoma de México
Menntavegur 1, Reykjavík, Z.P. 101, Iceland, Ciudad Universitaria, Mexico City, Mexico, CP. 04510
carlosm@ru.is,ivanvladimir@turing.iimas.unam.mx

Abstract

This work presents the path toward the creation of eight Spoken Language Resources under the umbrella of the Mexican *Social Service* national program. This program asks undergraduate students to donate time and work for the benefit of their society as a requirement to receive their degree. The program has thousands of options for the students who enroll. We show how we created a program which has resulted in the creation of open language resources which now are freely available in different repositories. We estimate that this exercise is equivalent to a budget of more than half a million US dollars. However, since the program is based on retribution from the students to their communities there has not been a necessity of a financial budget.

1. Introduction

In recent times there has been a rise in the number of available Language Resources for different speech processing and NLP tasks (Ray et al., 2018). However, this rise has not been equal for all languages and their variants (Hernández-Mena et al., 2017). The environment for the creation of Language Resources is different among regions and countries. In particular, for Latin America there has been reported a notable gap in the availability of resources among other aspects (Poblete and Pérez, 2020; Sanchez-Pi et al., 2022). In our experience, one of the main obstacles to the creation of resources this region is related to the economy, as research and industry budgets are small. Additionally, with recent ethics recommendations for fair pay, the creation of resources becomes more difficult, although, it is important to notice that fair payment is a necessity in several regions (Shmueli et al., 2021).

In this work, we present our approach to creating Language Resources for the Mexican and Latin American Spanish variants. To tackle the lack of financial resources, we rely on a national and institutionalized social program that every undergraduate has to comply with. This program is known in the region as *social service/Servicio Social*[1] which requires by law that an undergraduate student has to donate 480 hours of activities beneficial to society. In particular, this program encourages students to donate work hours in activities related to their field of study. By creating an option for students of engineering and linguistics, we have been able to collect up to 10 hours of speech per student, which has yielded eight freely available resources that support research in Mexican Spanish.

2. Context of Mexican *Social service*

Social service programs have been implemented around the world (UNESCO, 1984). In Mexico, the *Social Service* was started in 1935 as a requirement to be able to obtain an undergraduate degree in the *Universidad Nacional Autónoma de México* (UNAM). At that time it was the only institution that had such a requirement. By 1945 *Social Service* became a national program which any undergraduate student had to do by law. The main goal of the program was to give back to the society which had financed public education in Mexico and to allow students to acquire experience and practice in their field. In order to reach these goals, students have to apply to a registered option which is associated with a public institution. Once students join they have to donate 480 hours at a maximum rate of 20 hours per week (half time job), which guarantees that they spend at least six months in activities in support of society with a maximum duration of two years. Students have a great variety of options to enroll in. For most of the registered options, the students do not receive financial compensation; however, a few options requiring relocation can provide a scholarship.

3. Design of a *Social service* for Language Resources Creation

In 2013, one of the authors of this work established a *Social service* option for Engineering and Linguistics students at the UNAM's Engineering and Philosophy and Linguistics Faculties. The option was called "Development of Speech Tecnologies" and had the goal of creating speech resources and tools. This option was available for students until 2020 with a gap during 2015.

Since the start of the *Social Service* option, it was clear that this was a good opportunity to focus on activities that facilitated the creation of Language Resources, in particular for speech since there were not many resources that were open and freely available for research

[1] We use this translation for the name of the program in the absence of a better option. It is not related to the English terms *social services* or *social work*, but may be considered more similar to *community service* or *civic service*.

Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People @LREC2022, pages 20–24
Marseille, 25 June 2022.

or development. With this in mind, in the registered program students could perform any of the following activities:

Segmentation of audio: In this task students identify utterance segments in long audio recordings. These segments are fine-grained in the sense that they tend to be short (never shorter than 3 seconds). Students were provided with 30 hours of raw recordings and the goal was to have these hours segmented by the end of the *Social Service* commitment period. The recordings could come from sources such as radio-podcasts, talks or readings from books or Wikipedia articles. This task was performed using the *Audacity* software[2]. The software was chosen because it is open source and it was available in different operating systems and platforms.

Speaker-based segmentation: In this case, students identify sequences and segments composed of consecutive turns in which a single speaker speaks. For this task, students were provided with 50 hours of raw recordings to be segmented during the duration of their commitment period. This task was done using the *Audacity* software[3] as well.

Fine-grained speaker segmentation: Based on the segments from the previous task, students refine utterance segments. Since the segments are from a single speaker this is faster than the *Segmentation of audio* task that was done directly on the original recording[4].

Transcription of audio: Students orthographically transcribe what is said in utterance segments. During this task, students were asked to identify errors such as if the recording did not contain speech but another type of sound (e.g., music, background noises, etc). For this task the recommendation was to use the *Notepad++* software[5] which is easy to install and simple enough for the task.

The first year that the option was running, the pipeline consisted of two tasks: *Segmentation* and *transcription of audio*. However, the segmentations produced were not acceptable as they had a large amount of mistakes. The task was harder than originally planned; a student performing the first task of the pipeline will make mistakes at a higher rate than expected. After detecting this large number of errors, the segmentation task was split into two, so the pipeline consisted of three tasks: *Speaker-based segmentation*, *Fine-grained speaker segmentation* and *Transcription of audio*. We discovered there was a better coupling among the segmentations from the new first task and the new second task, since errors in the first task could be detected and fixed during the second one. The complexity of both tasks was less than the original approach because students do not have to worry about the length of segments as they cut, or worry about the order or the content of the recording; they just focus on the quality of the segment and speech.

For the segmentation task, it was important that students had a clear expectation of how the final audio segments should sound. To clarify this, the concept of *clean speech audio* was introduced with the following characteristics:

- There is only one speaker in the segment.
- There shouldn't be music on the background.
- The background noise should be minimal.
- There shouldn't be other types of human-produced sounds such as laughter or applause.

The *Social Service* option started with 3 students but by 2018 there were on average 60 enrolled students per year. The students did not receive any scholarship compensation for their service. However, we believe the popularity of the program derives from the following aspects:

1. The tasks could be performed at home. Although today we are very familiarized with the home-office modality of working, this characteristic was a novelty at the beginning of this option and it soon became very popular among the students. This option was a rarity compared to other options where they could do their *Servicio Social*. This was advantageous for students who lived far from the University or who had a limited amount of time (e.g., they worked to help their families or be able to pay for their studies).

2. The tasks could be done self-paced. At the beginning of the process, the students received a set of recordings that they could work on as it was convenient for them. They could decide their weekly load and schedule and adapt it depending on their availability.

To guarantee a homogeneous quality of the segmentations and transcriptions, students were provided with detailed manuals and some videos that explained the process at the conceptual level and illustrated its stages using the specialized software tools. Beside the instructions on the characteristics of the speech audio, the manuals include instructions about the naming of the files and their ordering in the corresponding folders. Of particular interest was to separate the recordings by

[2] Audacity audio editor website `https://www.audacityteam.org/` (last visited April 2022.)

[3] Idem.

[4] Idem.

[5] Notepad++ editor website: `https://notepad-plus-plus.org/` (last visited April 2022.)

two genders, male and female, and to try to be consistent with the speakers' identities, although the specific identities were discarded in the final version.

In the case of the segmentation task, the manual indicated the desired characteristics of the resulting audios:

- The audio must start and end with a small silence.
- The audio file should have the following format: Microsoft WAV, PCM, 16 bit signed.
- It should be mono (one channel).
- It should have a sampling rate of 16 kHz.
- The filename must include just ASCII characters with underscores between words instead of spaces.

In the case of the transcription task, the students were provided with the following requirements:

- Everything is transcribed in lower case.
- Numbers are transcribed orthographically, not using digits.
- Punctuation marks are not necessary.
- Mispronunciations are recorded in the spelling.
- Foreign words are transcribed as they sound, not with their native spelling.
- Acronyms also are transcribed as they sound.
- Alternative spellings should be avoid, particularly for not well known spellings with double letters, e.g. *clarissa*.
- In case of stuttering register the enunciation of it as much as possible.
- Disfluencies should be registered as sounding, in a short manner and capturing the vocal sound, e.g. *mmm* should be transcribed as *um*, *shhh* as *shu*, etc.
- Novel words should be registered and in case of accentuation (common in Spanish) this should be marked with the acute symbol.

4. Collected resources

There were eight Speech Resources, consisting of ten corpora, created through the *Social Service* option described in this work. All together they consist of 215 hours of speech. The difference between a Speech Resource and corpus is in their publication status; in particular, one Speech Resource could include more than one corpus, as will be shown in one of our speech resources. Table 1 shows the names of the Speech Resources, their size in hours, the year of publication and the repository where they are located. All but one of the resources were published at the Linguistic Data Consortium[6] (LDC) and the other at Open Speech and Language Resources [7] (OSLR).

Seven corpora were developed as part of the CIEMPIESS-UNAM project which was started to create the CIEMPIESS Corpus (Hernández-Mena and Herrera, 2015). The goal was to have a spontaneous speech corpus. It consists of recordings from 43 episodes of broadcast by Radio IUS, a UNAM radio station, with each episode being one hour long. Episodes are comprised of spontaneous conversations between a radio moderator and guests, and their main topic is legal issues. Approximately 78% of the speakers were males, and the rest were females. At a later time, **CIEMPIES Light** (Hernández-Mena and Herrera, 2017) was released, which was an updated and improved CIEMPIESS version but it did not include the automatic phonological transcriptions that the original resources did. This corpus also was designed to be easy to use with Kaldi software (Povey et al., 2011).

A problem with the CIEMPIESS and CIEMPIESS Light corpora was that they are unbalanced, particularly because there are few female speakers. In order to solve this bias, two new resources were created: **CIEMPIESS Balance** (Hernández-Mena, 2018) and **CIEMPIESS experimentation** (Hernández-Mena, 2019a). The first one is the inverse image of the CIEMPIESS corpus (Hernández-Mena and Herrera, 2015) in terms of gender since it contains more speech from female speakers than male. Its goal was that once combined with CIEMPIESS Light, both would produce a gender balanced corpus. On the other hand, the CIEMPEISS Experimentation resource consists of three corpora: *Complementary*, *Fem* and *Test*. These corpora had a specific goal: the *Complementary* corpus consists of a minimal set of utterances to constitute a phonetically balanced corpus; *Fem* consists of the remaining transcriptions of female speakers' recordings that were not included in Balanced; finally, *Test* is a test set of spontaneous speech.

It was during the beginning of the CIEMPIESS-UNAM project that the *Social Service* contributed to the creation of the **CHM150** corpus (Hernández-Mena and Herrera, 2016). This corpus is comprised of Mexican Spanish microphone speech from 75 male and 75 female speakers in a quiet office environment. The speech is spontaneous, triggered by open questions or by requesting the description of a painting shown to the speaker on a computer monitor. Its characteristics make it a candidate to be an evaluation corpus, but it is a challenging corpus since the speech is spontaneous.

Since the series of resources associated with the CIEMPIESS project was spontaneous speech, there

[6] LDC website: `https://catalog.ldc.upenn.edu/` (last visited April 2022.)

[7] OSLR website: `https://openslr.org` (last visited April 2022.)

was an additional effort to create resources around read speech. For this it was decided to use LibriVox[8] (Hernández-Mena, 2020) which collects open and freely available readings of public domain books, and *Wikipedia grabada*[9] (Hernández-Mena and Ruiz, 2021) which is composed of reading recordings from Wikipedia articles.

Finally, the team decided to work on the TEDx collection of talks. For this a new corpus was proposed (Hernández-Mena, 2019b). The speech in this resource is spontaneous; however, there are large monologues which helped with the segmentation of it and to process it in a timely fashion.

Corpus	Size	Published
CIEMPIESS	17h	LDC/2015
CHM150	1.6h	LDC/2016
CIEMPIESS Light	18h	LDC/2017
CIEMPIESS Balance	18h	LDC/2018
CIEMPIESS Experimentation	40h	LDC/2019
TEDx Spanish	24h	OSLR/2019
LibriVox Spanish	73h	LDC/2020
Wikipedia Spanish	25h	LDC/2021

Table 1: Corpora produced by the *Social Service* option described in this work, size given in hours.

5. Ethical concerns

We are aware that there could be concerns that not paying the students is not a fair situation. In fact most of the *Social service* options in Mexico are without payment, and from the legal point of view the law gives that prerogative to the administrator of the *Social Service* option. From the social and ethical point of view, the implicit contract in Mexican society is that students have to give back, particularly in the public system in which students receive a free education. The program is based on a reciprocation principle. In our case the program here described had the goal of not becoming an exploitation case where students work more than what the laws require. To achieve this we implemented the following policies:

- The work load was calculated for 480 hours, and it was constantly validated for the different tasks.

- As mentioned, we provided a maximum flexibility to perform the assigned task. For example, some students did not finish as planned when the 2020 COVID-19 pandemic started, so together with the schools we allowed them to finish their process from one to up to two years after.

- Minimum hardware requirements: the chosen software guaranteed that required computer power was minimal, and no specific brand or OS was necessary. Students had a heterogeneous set of computers and this flexibility allowed them to use their current machines for the work. Also, we did not use online based software since many of them had restricted Internet access.

- To guarantee the impact of the students' work, the created resources were released under an open and freely available license. This was explained to the students at the beginning of the commitment period.

What has to be highlighted about these resources is that there might exist some implicit bias in the work since the segmenter and transcriber population is comprised of undergraduate students. This is something to have in mind since it is the population that the social service program is addressed toward.

6. Conclusion

In this work we describe the use of a *Social Service* option to create Language Resources, in particular for Speech. From our calculations we estimate that the 480 hours invested in this project corresponds approximately to $4,000$ USD, which amounts to an investment of $800,000$ USD when we consider that more than 200 students have enrolled and contributed to the creation of freely and openly available resources. For us, this exemplifies a success story of this approach in which solidarity and retribution from the students allow the collection of large resources of Spoken Mexican Spanish.

As future work we plan to continue working with speech resources, since we need more resources to capture well the richness of the region, particularly for spontaneous speech and multiple speaker scenarios. However, we would like to explore large collaborations based on extended reciprocation principles. For instance, we would like to collaborate with public institutions in which social service is not established but can provide training or scholarships to the students to continue to develop open language resources.

7. Acknowledgements

The authors thank the students of the Social Service option ”Desarrollo de Tecnologías del Habla” from the Facultad de Ingeniería (FI) and Facultad de Filosofía y Letras (FFyL) of the Universidad Nacional Autónoma de México (UNAM), we are thankful for all the hard work. The authors also thank Caitlin Laura Richter for her suggestions and comments on the current version of the paper. Carlos Hernández-Mena thanks the support from Language and Voice Lab from Reykjavik University in the realization of this manuscript.

[8] LibriVox website `https://librivox.org/` (last visited April 2022)

[9] Wikipedia *grabada* website `https://es.wikipedia.org/wiki/Wikiproyecto:Wikipedia_grabada` (last visited April 2022)

8. Bibliographical References

Hernández-Mena, C. D., Meza-Ruiz, I. V., and Herrera, A. (2017). Automatic speech recognizers for mexican spanish and its open resources. *Journal of applied research and technology*, 15(3):259–270.

Poblete, B. and Pérez, J. (2020). Minding the ai gap in latam. *Communications of the ACM*, 63(11):61–63.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Ray, J., Johnny, O., Trovati, M., Sotiriadis, S., and Bessis, N. (2018). The rise of big data science: A survey of techniques, methods and approaches in the field of natural language processing and network theory. *Big Data and Cognitive Computing*, 2(3):22.

Sanchez-Pi, N., Martí, L., Garcia, A. B., Yates, R., Vellasco, M., and Coello, C. (2022). A roadmap for ai in latin america. *PLoS neglected tropical diseases*.

Shmueli, B., Fell, J., Ray, S., and Ku, L.-W. (2021). Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.

UNESCO. (1984). El servicio social universitario un instrumento de innovaciÓn en la educaciÓn superior. Technical report, UNESCO.

9. Language Resource References

Carlos Daniel Hernández-Mena and Abel Herrera. (2015). *CIEMPIESS LDC2015S07*. Linguistic Data Consortium, CIEMPIESS-UNAM project, ISLRN 838-468-581-053-6.

Carlos Daniel Hernández-Mena and Abel Herrera. (2016). *CHM150 LDC2016S04*. Linguistic Data Consortium, CIEMPIESS-UNAM project, ISLRN 649-160-209-726-6.

Carlos Daniel Hernández-Mena and Abel Herrera. (2017). *CIEMPIESS Light LDC2017S23*. Linguistic Data Consortium, CIEMPIESS-UNAM project, ISLRN 273-364-546-427-6.

Carlos Daniel Hernández-Mena and Iván Vladimir Meza Ruiz. (2021). *Wikipedia Spanish Speech and Transcripts LDC2021S07*. Linguistic Data Consortium, ISLRN 676-370-775-701-9.

Carlos Daniel Hernández-Mena. (2018). *CIEMPIESS Balance LDC2018S11*. Linguistic Data Consortium, CIEMPIESS-UNAM project, ISLRN 304-456-056-609-5.

Carlos Daniel Hernández-Mena. (2019a). *CIEMPIESS Experimentation LDC2019S07*. Linguistic Data Consortium, CIEMPIESS-UNAM project, ISLRN 139-696-537-175-5.

Carlos Daniel Hernández-Mena. (2019b). *TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license*.

Carlos Daniel Hernández-Mena. (2020). *LibriVox Spanish LDC2020S01*. Linguistic Data Consortium, ISLRN 256-321-086-598-2.