From Field Linguistics to NLP: Creating a curated dataset in Amuzgo language

Antonio Reyes Pérez
Facultad de Lenguas y Letras, UAQ
antonio.reyesp@uaq.edu.mx

Hamlet Antonio García Zuñiga
Instituto Nacional de Antropología e Historia
hamlet_garcia@inah.gob.mx

Abstract

This article presents an ongoing research on one of the several native languages of the Americas: Amuzgo or $jny'on^3\ nda^3$. This language is spoken in Southern Mexico and belongs to the Otomanguean family. Although Amuzgo vitality is stable and there are some available resources, such as grammars, dictionaries, or literature, its digital inclusion is emerging (cf. Eberhard et al. (2024)). In this respect, here is described the creation of a curated dataset in Amuzgo. This resource is intended to contribute to the development of tools for scarce resources languages by providing fine-grained linguistic information in different layers: From data collection with native speakers to data annotation. The dataset was built according to the following method: i) data collection in Amuzgo by means of linguistic fieldwork; ii) acoustic data processing; iii) data transcription; iv) glossing and translating data into Spanish; v) semiautomatic alignment of translations; and vi) data systematization. This resource is released as an open access dataset to foster the academic community to explore the richness of this language.

1 Introduction

According to the facts reported in the survey *Analysis of the Language Technologies in Mexico* (ASTLM, 2018), Latin America is a linguistic region with a minimum development in the creation of digital resources, specifically, regarding native languages. One of the main causes, pointed out by the authors, is the scarcity of data. The researchers frequently face a lack of materials to study these languages. For instance, several native languages have not even been described, either because they have not been considered academically, or because their grammar is difficult, or data collection is complicated or, even, highly risky. This stresses a gradual loss of worldviews, as well as an augment of the digital divide, which will directly impact on the native communities by making inequalities and marginalization larger.

With respect to the Mexican context, specialized organisms such as the National Institute of Statistics and Geography or the National Institute of Indigenous Languages report an important linguistic wealth. According to the numbers registered in (INALI, 2008), apart from Spanish, more than 60 native languages coexist in the country. This diversity is classified in 11 linguistic families, 68 languages, and 364 dialectal varieties. In terms of their vitality, the languages with more speakers are Nahuatl (more than one million), Mayan (around 800,000), Mixtec, and Zapotec (over 400,000 speakers each). On the contrary, there exist an important number of languages with fewer than 1,000 speakers (INEGI, (2015). Sadly, despite this enormous linguistic diversity, there is still a lack of resources, tools, and even linguistic materials for the majority of these languages (surprisingly, some of them well studied and described).

Amuzgo or $jny'on^3\ nda^3$ is one of the languages cited in the previous reports. Currently, the language has a couple of grammars and several studies about its varieties (see (Buck, 2018; Smith and Tapia, 2002; Palancar and Feist, 2015; Hernández et al., 2017)). In accordance with these specialized works, Amuzgo is quite different from the major language in the country (Spanish), which makes it more complex to directly apply techniques or tools from other languages. In this respect, this work describes an ongoing research about the creation of a dataset in Amuzgo. The dataset contains curated linguistic data collected from colloquial speech in Amuzgo. These data are presented considering the following phases or levels: acoustic signal processing, transcription, glossed and human translation into Spanish, semiautomatic alignment of human translations, and annotation. This dataset is intended to contribute the development of tools for scarce resources languages.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 127–131

The rest of the article is organized as follows: Section 2 describes the main linguistic characteristics of Amuzgo, focusing on their complexity. Section 3 presents the method to create the dataset. Section 4 points out some of the results and applications so far. Finally, Section 5 summarizes our findings and highlights the future work.

2 Language description

Amuzgo or $jny'on^3\ nda^3$ is spoken in Southern Mexico. It has over 60,000 speakers, according to the data reported by (INEGI, (2015). Grammatically, the language belongs to the Otomanguean family, alongside languages such as chatino, zapotec, mazatec, or popolaca. It is characterized by a wide set of personal pronouns (Buck, 2000; Palancar and Feist, 2015), and lexical classes, which impacts on its verbal complexity (Smith and Tapia, 2002). In addition, this is a language with five tones to mark change of meaning in different levels (Hernández, 2019). For instance, Examples (1) and (2) show how the different tones impact, not only on the meaning, but in the linguistic level as well.

(1) a. $nkia^3$ "to hit"
b. $nkia^5$ "fearful"

(2) a. ba'^1 "his/her house"
b. ba'^4 "your house"

While in (1), the difference between the low level tone and the high level tone (superscripts 3 and 5, respectively) impacts on a change in the lexical level; i.e. the tonal contrast produces two different types of words, in (2), the change moves to the morphological level; i.e. the tonal change (rising tone, superscript 1, and mid level tone, superscript 4) generates a clear distinction with respect to who possesses the house.

2.1 Linguistic complexity

This couple of examples stress how complex the language is, and somehow, how difficult is to apply tools or techniques that have shown their usefulness in other languages. Now, to summarize the challenges that this language entails, both in terms of its linguistic description and its possible digital implementation, we provide below some of the mos salient linguistic particularities reported in the specialized literature.

The phonological system in Amuzgo has 15 consonants and 7 vowels. Some of these phonemes are product of the contact with Spanish. As noted previously, its tonal characteristics impact on different linguistic levels: From phonology to pragmatics. In this respect, it has been reported the existence of five tones: rising, falling, low level, mid level, and high level. However, some works report two more tones; i.e. five level tones and two contour tones (Hernández, 2019).

On the other hand, Amuzgo organizes the grammatical persons in two classes: Singular and plural. Each class contains three persons morphologically marked, being the particular interest the third person in plural (they), which differentiates between inclusion and exclusion regarding the hearer.

Finally, in terms of morphosyntactic and syntactic particularities, Amuzgo is a head-marking language, but for the third persons; the syntactic relations are given by juxtaposition; and the syntactic template both for transitive and intransitive verbs is VSO.

3 Method

In this section, we describe the method to build the dataset. Specifically, we highlight how the dataset is curated by means of adding fine-grained linguistic information to the gathered data. The method consist of the following phases: i) collecting acoustic data from fieldwork, ii) acoustic data processing with Praat, iii) data transcription, iv) data glossed and human translation, v) semiautomatic alignment of human translations, and vi) data systematization.

3.1 Data collection

The data was collected by means of traditional linguistic fieldwork in one of the Amuzgo communities. Some of the researchers traveled to the community to interview a set of speakers in their natural environment. The Amuzgo speakers were asked some questions about how frequent they speak the language, the contexts in which they use it, their linguistic skills, both in Amuzgo and Spanish, whether or not they consider themselves monolingual or bilingual speakers, as well as information about their age, gender, time of residence in the community, whether or not they were migrants, and so on. In addition, they had to tell a story in Amuzgo. All the stories were digitally recorded. The amount of hours of this oral material is around 8 hours. Finally, it is important to remark that all the participants were informed about the use of the material in accordance with the Mexican laws. If

they agreed, then they should sign a document. A more detailed explanation about this process is provided in the project GitHub site (see Section 3.5).

3.2 Acoustic data processing

The dataset here presented is based on the 8 hours of recordings. This oral material is, thus, the raw data from which we build the dataset.

Prior to adding any linguistic annotation, the oral material was acoustically processed with the software Praat (Boersma, 2014). This was done in order to get unbiased phonological information to be further analyzed, especially, given the tonal characteristics of the language reported in Section 2.

It is worth stressing that the acoustic signal processing is work in progress. So far, only three recordings have been fully processed. Once this phase is complete, it will be a major outcome of the research.

3.3 Data transcription

The following phase consisted in making the transcription of the oral data. Firstly, the stories were divided in segments to facilitate the transcription; i.e. we did not transcribe each story at once; on the contrary, we attempted to recognize informative segments or sentences to achieve a coherent text. In this recognition process, the work performed with the tools for speech analysis in Praat were highly helpful, due to they were a visual guide to segment the recordings properly.

The resulting text was linguistically analyzed in order to corroborate the initial segmentation. In case of unnatural divisions, for instance, incomplete sentences, the transcription was modified to adapt as much as possible each segment with a clause. This decision impacts directly on the glosses and the human translation, yet it did not affect the language grammaticality.

The last process of the transcription consisted in registering the Amuzgo phonological features, as well as incorporating the tones. The following text is a sample of the resulting transcription.

Twe'^{3} nkwi^{3}xue^{12} m'an^{3} kwi^{3}ti'^{12}tyo^{3}ndye^{35} ts'a^{3} ti'^{12},
ts'ian^{5} jndë^{12}, tyua'^{12} ju'^{35} sku'^{35} ti'^{12} k'a^{35} ti'^{12} jndë^{12},
Mo'^{12} twe'^{3} nkwi^{3} xue^{5} t-ja^{3} ti'^{12},
të^{1}ki^{3}tsa^{35} ti'^{12} ts'ian^{5},
no'^{1} ya^{12} tje^{35} ti'^{12} tyua'^{12}je^{12},

3.4 Glosses and human translation

In these two processes, glosses and human translation, rely the major richness of the dataset. With respect to the former, it is well-known that a glossed process entails a syntactic segmentation (just like the one attempted in the previous phase), as well as a part of speech annotation. This linguistic information contributes to enhance the description and understanding of any language, due it provides an in-depth vision of its linguistic relations. In addition, the glosses provide formal elements that can be used to perform a better translation.

The glosses were done manually, i.e. a human expert in Amuzgo analyzed the transcriptions, made a re-alignment of the unnatural segments proposed in the previous phase, and performed the glossed process according to the Leipzig Glossing Rules (Comrie et al., 2008). Furthermore, the expert considered the following elements when generating the glosses: use of a consistent orthographic system (Hernández, 2019), distinction between phonological word and lexical entry (see Examples (1) and (2)), and clitic marking.

The following process was a human translation, which was guided by the linguistic information registered in the glosses. Figure 1 illustrates the result of the glossing and human translation.

3.5 Translation alignment

This phase was performed to enhance the scope of the dataset. Mainly, by creating a parallel corpus Amuzgo-Spanish, which was generated semi-automatically by applying the Gale-Church algorithm for translation alignment (Gale and Church, 1993). This algorithm is implemented in the CAT tool OmegaT and it considers two methods: Parsewise and heapwise. Each one makes the alignment taking into consideration different features in the texts. For instance, a possible syntactic parallelism regarding the parsewise method, or a global textual integration regarding the heapwise method. We tried with both methods and the result was unsatisfactory. This is obviously due to the huge linguistic differences between both languages.

Despite the poor performance, the results were used as a source to manually improve the alignments. This human improvement produced a better parallel corpus, which is freely available in this address: https://github.com/areyesp-77/amuzgo-dataset.git.

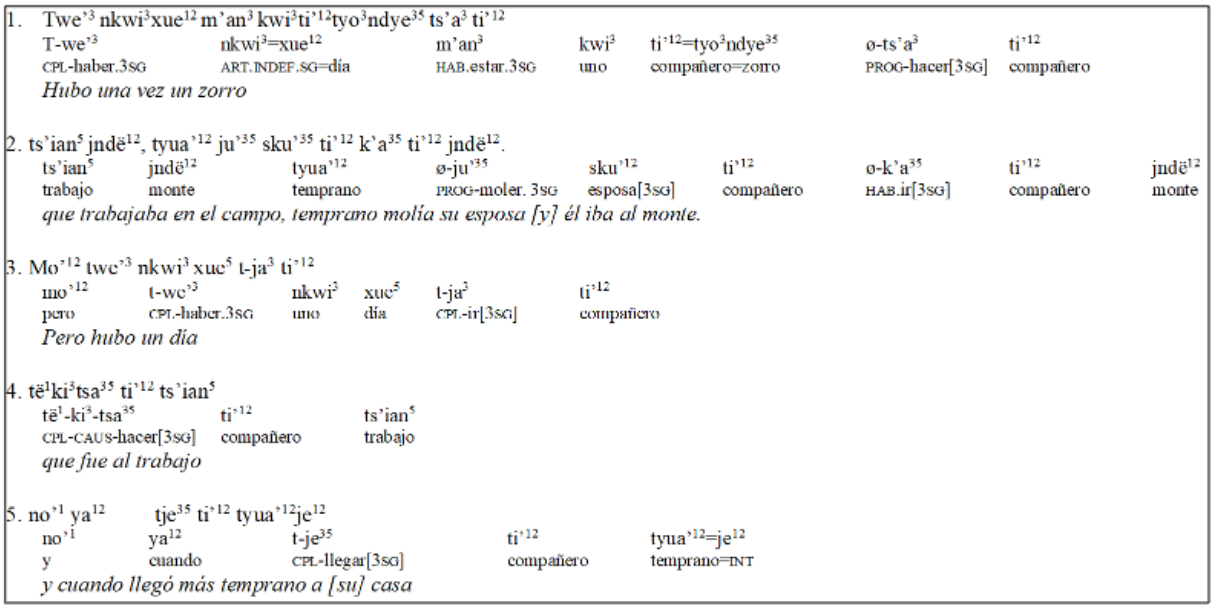

Figure 1: Sample of the glosses and human translation.

3.6 Data systematization

The last phase to create the dataset implies the integration of the linguistic information in a lexical resource to expand the possibilities of investigation, as well as to start applying some NLP techniques in tasks as diverse as machine translation, part of speech tagging, speech recognition, and so on. To this end, we have been working in systematizing and formalizing our data in a unique file to simplify, as much as possible, all the fine-grained linguistic information that we have.

In this respect, a beta version of this curated dataset with the following information has been released: lexical entry in Amuzgo, POS annotation, tone annotation, linguistic processes identified, translation into Spanish, and source (recording). This version is available at: www.geco.unam.mx.

4 Result and projection

The outcome of all the processes described so far is a dataset with fine-grained linguistic information in Amuzgo. Beyond the traditional linguistic analysis, this information could be used to foster and increase the efforts of the NLP community regarding scarce resources languages, such as $jny'on^3$ nda^3.

In addition, a small parallel corpus Amuzgo-Spanish has been generated to freely explore some characteristics of the language by using common corpus analyses, such as concordances, collocations, or keywords.

As an optimistic projection, the creation of resources like this dataset would allow the experimentation in different areas, the generation of new knowledge, the preservation of endangered languages, and the minimization of the digital divide and its negative consequences for the native communities.

5 Conclusions

In this article we have described the creation of a linguistic dataset in Amuzgo. The dataset contains curated information provided by specialists in that language. Such information covers different linguistic levels, such phonology, morphology, and syntax, as well as a human translation into Spanish. In addition, the dataset was used to generate a small parallel corpus Amuzgo-Spanish by applying a known algorithm in NLP. Both resources are freely available for academic purposes.

Although the materials exceed the 8 hours of recordings, the dataset here described contains only the data of one hour. Therefore, as further work, it is planned to process the remaining data to enhance the linguistic description and, accordingly, to improve the dataset.

References

ASTLM. 2018. *Análisis del sector de las Tecnologías del lenguaje en México*. Gobierno de España.

Paul Boersma. 2014. The Use of Praat in Corpus Research. In Ulrike Gut Jacques Durand, and

Gjert Kristoffersen, editor, *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

M. Buck. 2000. *Gramática del amuzgo de San Pedro Amuzgos*. Instituto Lingüístico de Verano.

M. Buck. 2018. *Gramática del amuzgo de Xochistlahuaca*. Instituto Lingüístico de Verano.

B. Comrie, M. Haspelmath, and B. Bickel. 2008. *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses*. Max Planck Institute for Evolutional Anthropology.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2024. *Ethnologue: Languages of the World*. SIL International.

W. Gale and K. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

N. Hernández. 2019. *El sistema tonal en el amuzgo de San Pedro Amuzgos: Interacción entre el tono de la base nominal y los clíticos*. Tesis de Maestría en Lingüística Indoamericana, Ciudad de México.

N. Hernández, A. Mora, and H. García. 2017. Estructura de la frase nominal posesiva en amuzgo (otomangue). *UniverSOS. Revista de Lenguas Indígenas y Universos Culturales*, 14:63–82.

INALI. 2008. Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas. *En Diario Oficial de la Federación*.

INEGI. (2015). Encuesta intercensal 2015. url =.

E. Palancar and T. Feist. 2015. Agreeing with subjects in number: The rare Split of Amuzgo verbal inflection. *Linguistic Typology*, 93(3):337–383.

T. Smith and F. Tapia. 2002. Amuzgo como lengua activa. pages 81–129. Del cora al maya yucateco. Estudios lingüísticos sobre algunas lenguas indígenas mexicanas.