Collaborative Corpus Creation: A Ch'ol Case Study

Carol Rose Little, Juan Jesús Vázquez Álvarez, Jessica Coon, Nicolás Arcos López, and Morelia Vázquez Martínez

1. Introduction

The aim of this paper is to describe two collaborative linguistic research and documentation projects, which culminated in two corpora of Ch'ol narratives, available through the Archive of Indigenous Languages of Latin America (AILLA). We outline how the projects served the twin goals of (i) facilitating linguistic research and documentation materials on Ch'ol, and (ii) increasing language awareness and building capacity among Ch'ol-speaking students through the process of corpus creation. We discuss how this "crowd-sourcing" approach to linguistic corpus creation has the potential to benefit both language communities and researchers. The authors of this paper represent the different roles of participants in the project: native-speaker linguists working in universities in southern Mexico, linguists in the US and Canada, as well as a Ch'ol-speaking student.

The first project titled *Documenting word order variation in Mayan languages: A collection of Ch'ol narratives* took place in 2018 and was funded by a National Geographic Society Explorers Grant awarded to authors Jessica Coon and Juan Jesús Vázquez Álvarez. Vázquez Álvarez is a linguist at Center for Multidisciplinary Research about Chiapas and the Southern Border (CIMSUR-UNAM) in San Cristóbal de las Casas, Chiapas, Mexico. His MA and PhD theses have been the basis for much descriptive and theoretical work on Ch'ol. Vázquez Álvarez is also a native speaker of the Tila dialect of Ch'ol. Jessica Coon is a linguist at McGill University, with a specialization in Mayan languages. In the National Geographic project, authors Nicolás Arcos López and Morelia Vázquez Martínez also participated. Arcos López is a professor in the Department of Languages and Cultures at the Intercultural University of the State of Tabasco (UIET) and also a native speaker of Ch'ol. Vázquez Martínez worked as a research assistant for the project and is a speaker of Ch'ol.

The second corpus was created for author Carol Rose Little's PhD thesis. The grant titled *Investigating Overt Definite Articles and Grammatical Variation* was funded by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation. This multidialectal corpus was used for investigating definiteness across dialects of Ch'ol. Author Vázquez Martínez also worked in the collection, transcription, translation of the corpus as well as the subsequent research projects that drew on material from this corpus.

In the following sections, we first provide background on the linguistic context of Ch'ol (Section 2). We then detail the creation of the two corpora in Section 3. We discuss the process of creating the first corpus in 3.1 and how the stages of the creation of the first corpus led to capacity-building for Ch'ol-speaking university students. In 3.2, we discuss the creation of the second corpus for Little's dissertation research. Section 4 presents how the corpora have been used for research and resources in Ch'ol as well

* Carol Rose Little, McGill University (crl223@cornell.edu), Juan Jesús Vázquez Álvarez, CIMSUR-UNAM, (juanvazquezalvarez@gmail.com) (corresponding authors), Jessica Coon, McGill University, Nicolás Arcos López, Intercultural University of the State of Tabasco (UIET), Morelia Vázquez Martínez. We would like to thank students and professors at UIET and the Multidisciplinary Academic Unity (UAM) at the Intercultural University of Chiapas (UNICH) in Yajalón. Special thanks go to Prof. Bernabé Vázquez at UAM-UNICH in Yajalón for his role in organizing workshops and setting up the students so that they could get university credit for their work in this project. Funding for these projects thanks to a National Geographic Society Explorers Grant, "Documenting word order variation in Mayan languages: A collection of Ch'ol narratives" awarded to Jessica Coon and Juan Jesús Vázquez Álvarez. The second corpus is based upon work supported by the National Science Foundation under grant no. BCS-1852744 and an Engaged Cornell graduate student research grant awarded to Carol Rose Little.

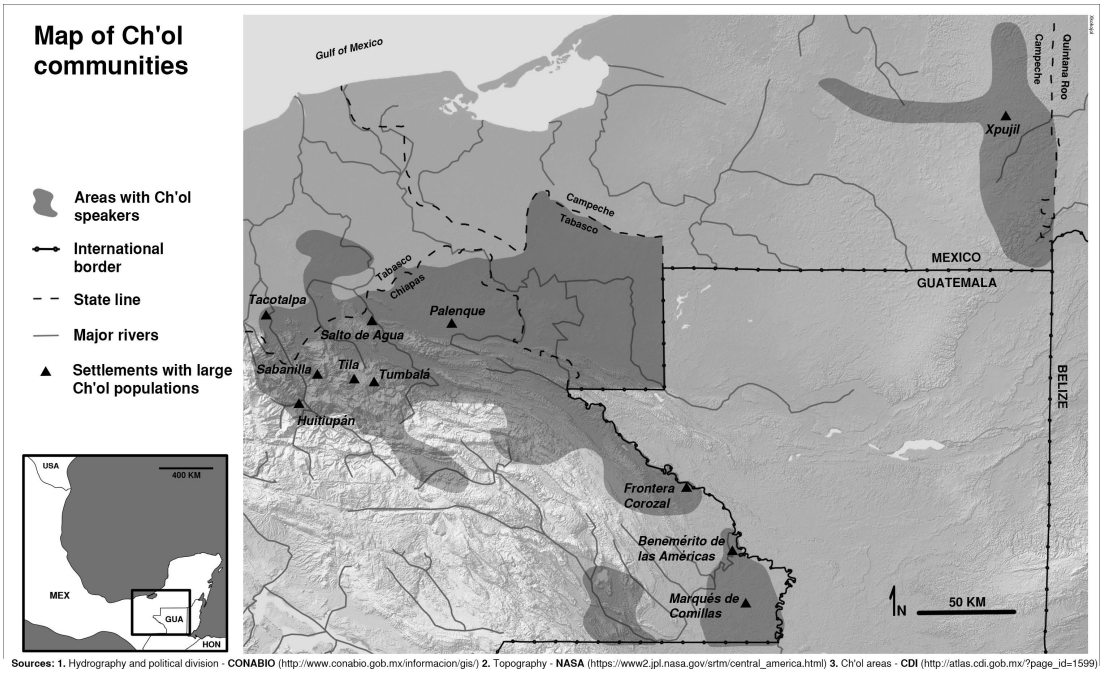
 Proceedings of the 42nd West Coast Conference on Formal Linguistics, ed. Shweta Akolkar et al., 478-487. Somerville, MA: Cascadilla Proceedings Project.

as extensions to pedagogical resources to promote metalinguistic awareness amongst Ch'ol-speakers. We also make note of a presentation given in Ch'ol on definiteness by Vázquez Martínez, highlighting the significance of conveying research about Ch'ol *in* Ch'ol. Section 5 concludes with a description of how a similar corpus-creation model has been applied to the Algonquian language Cheyenne in Montana.

2. Background on Ch'ol

Ch'ol is a Mayan language of the Ch'olan-Tseltalan branch spoken in Southern Mexico, as depicted in the map in Figure 1. Ethnologue (2020) reports there to be 252,000 speakers, citing 2015 statistics from INEGI (*Instituto Nacional de Estadística y Geografía*, "National Institute of Statistics and Geography" in Mexico). There are three mutually intelligible dialects of Ch'ol: Tumbalá, Tila and Sabanilla. These dialects differ slightly in their phonology and morphology, but the greatest variation is seen in their lexicon. For instance the progressive marker is *choñkol* in the Tila dialect, *woli* in Tumbalá and *yäkel* in Sabanilla. The Tila dialect is spoken in the municipalities of Tila and parts of Tabasco; the Tumbalá dialect is spoken in the municipalities of Tumbalá, Salto de Agua and Palenque; the Sabanilla dialect is spoken in Sabanilla and Huitiupán. As we will discuss below, all three dialects are represented in the corpora.

Figure 1: Map of Ch'ol communities (Little 2020: 10)



Ethnologue categorizes the language status of Ch'ol as 5 for "developing" meaning that "[t]he language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable" (Ethnologue 2021). While Ch'ol is still used by multiple generations, language shift towards Spanish has been observed, especially among younger generations of speakers. Gómez Jiménez (2015: 19) reports that Spanish is displacing Ch'ol in many contexts, resulting in a loss of both the Ch'ol language and culture. For example, Ch'ol is no longer used as the main language of communication in local governments. This is especially visible in younger generations of Ch'ol speakers as many harbor negative attitudes towards their language due to feelings of "marginalization, discrimination, and other conditions imposed by the Mexican state" (Gómez Jiménez 2015: 18). These factors therefore signify that the use of Ch'ol is not yet "widespread or sustainable."

Nevertheless, there are many ongoing efforts by Ch'ol teachers, linguists, authors and scholars to further foster the preservation and maintenance of the Ch'ol language and culture. These include those working at institutions such as the *Centro Estatal de Lenguas, Arte y Literatura Indígenas* (The Center for

Indigenous Languages, Art and Literature of Chiapas, CELALI), *Casas de la Cultura* (Cultural Houses) in municipalities such as Salto de Agua and Tumbalá, and *Universidades Interculturales* (intercultural universities) in Tabasco and Chiapas. Below we detail how the creation of the two corpora add to the growing efforts of communities to document and preserve their language.

3. The two Ch'ol corpora

In the following two subsections, we describe the processes of how each corpus was created, the content of each corpus and the positive reception each corpus received in the communities.

3.1. Corpus 1: A crowd-sourcing model for corpus creation

Three of the authors (Coon, Arcos López and Vázquez Álvarez) were involved in the National Geographic-funded project to create a corpus of Ch'ol narratives. Ch'ol-speaking students, primarily undergraduates, were trained in a series of workshops in all aspects of documentation, from recording and consent, to transcription and orthography. The resulting corpus presents the material that came out of these workshops, collected and curated by Ch'ol-speaking students in their home communities over the course of six months throughout Chiapas and Tabasco, states in southern Mexico. The types of stories and materials collected reflect the choices of the community members themselves, many of whom developed further interest in learning about linguistics and language documentation. Below we describe the stages involved in the creation of the corpus, archived in AILLA (Vázquez Álvarez & Coon 2019).

During the first stage, a group of 25 undergraduates at the Intercultural University of the State of Tabasco—primarily Ch'ol-speaking students completing a major in Language and Culture—were recruited for a workshop devoted to understanding the process of recording. After initial discussion of the larger goals of language documentation and the importance of linguistic diversity, the workshop was divided into three parts: (i) "before recording" (equipment setup, location, consent, metadata, genres of recording, topics to avoid); (ii) a hands-on practice session with recording equipment; and (iii) a discussion of the larger goals of the project, including a preview of the ELAN transcription tool for audio and video recordings as well as the AILLA corpus page. A similar workshop was held at the Multidisciplinary Academic Unit (UAM), part of the Intercultural University of Chiapas (UNICH), in the Ch'ol-speaking town of Yajalón. Fifteen students participated in this workshop, again primarily Ch'ol-speaking undergraduates enrolled in a Language and Culture major. Students received diplomas for their participation and some received course credit.

In the second part of the project, with local supervision and support from local Ch'ol-speaking professors, including Arcos López, students who participated in the first recording workshops took turns checking out audio recorders and heading to their home communities to record friends and family members on the weekends. As a result, the corpus contains material from a number of different Ch'ol communities, spanning the major dialect regions. Students backed up their recordings on a central hard-drive, and received compensation for their travel costs in the form of small stipends. Stipends, instead of reimbursements, were used due to the difficult nature of collecting receipts in rural parts of Chiapas, where many of the students were traveling.

The third part of the project consisted in a transcription workshop. Students who chose to continue from both Tabasco and Chiapas groups convened in San Cristóbal de las Casas for a two-day ELAN transcription and Ch'ol orthography and writing workshop. The event was hosted at CIESAS-Sureste, a research center in San Cristóbal de las Casas, and included training on ELAN, as well as discussion of problems encountered and uncertainties in the standard Ch'ol writing system. Students came prepared with a selection of their favorite recordings for a hands-on ELAN practice session.

After the workshop, professors and students continued to work on the transcription of recorded material collected during the first months of the project. Undergraduate students received financial support for the time they spent transcribing, and some received service credit at their universities for their participation in the project. In addition to the work on the new material, Ch'ol material from previous Ch'ol-speaking linguists was collected, curated, and added to the corpus. Ch'ol-speaking research assistants Matilde Vazquez Vázquez and Morelia Vázquez Martínez (also an author)—each with previous experience in recording and Ch'ol transcription—assisted in editing, transcribing, and obtaining consent for older material.

The bulk of the material was collected in 2018, though older recorded material from Vázquez Álvarez, as well as other Ch'ol researchers, was curated and included in the corpus. The narratives range from personal anecdotes, to traditional stories, legends, and historical events in the region. Many of the recordings have been transcribed and translated, and each includes metadata with the place and date of recording, the title and description in Ch'ol and English, the researcher collecting the material, and the names of the participants being recorded. The resulting corpus contains:

- 191 WAV audio recordings (totaling 39 hours 48 minutes)
- 56 MPG, MP4, and AVI video recordings (totaling 3 hours 23 minutes)
- 72 XML transcription files in the EAF format for use with the ELAN application
- 2 JPG photographs

Table 1 provides a selection of the recordings. As can be seen, the recordings addressed various themes from traditional stories, historical events, instructions, and personal stories. All dialects of Ch'ol are represented: Tumbalá, Tila, and Sabanilla.

Table 1: Examples of recordings in AILLA for the corpus Vázquez Álvarez & Coon (2019)

Recording title	Location recorded	Contributors
The earthquake *Yujkel*	Hidalgo Joshil, Tumbalá, Chiapas	Nilda Patricia Gúzman López (Recorder) Pedro Guzmán Mayo (Speaker)
Earthquake and grasshoppers *Yujkel yik'oty sajk'*	Tila, Chiapas	Juan Jesús Vázquez Álvarez (Interviewer, Recorder) Francisco Pérez (Speaker) Guadalupe Vázquez Vásquez (Speaker)
The envied younger child *Xuty mu'bä its'a'leñtyel*	Teoquipa el Bascan, Salto de Agua, Chiapas	Estela Álvaro Díaz (Recorder) Francisco Álvaro Solís (Speaker)
Nahual *Wäy*	Tocob Leglemal, Tila, Chiapas	Adrián Sánchez Méndez (Recorder) Francisco Méndez Arcos (Speaker)
Process for a milpa *Bajche mi lakmel lakchol*	Hidalgo Joshil, Tumbalá, Chiapas	Juan Mario Mayo Hidalgo (Recorder) Jorge López Mayo (Speaker)
Working in Tabasco and Campeche *Troñel tyi Tabasco yik'oty tyi Campeche*	Tila, Chiapas	Juan Jesús Vázquez Álvarez (Interviewer, Recorder) Vicente Ramírez Vázquez (Speaker)
The search for food in the past *Isäkläñtyel bälñäk'äl wajali*	Tila, Chiapas	Juan Jesús Vázquez Álvarez (Interviewer, Recorder) Vicente Ramírez Vázquez (Speaker)
The word of God *Ity'añ lakyum*	Hidalgo Joshil, Tumbalá, Chiapas	Nilda Patricia Guzmán López (Recorder) Carlos Peñate Guzmán (Speaker)
Zapatistas *Zapatistajob*	San Juan el Mirador, Sabanilla, Chiapas	Félix Ignacio López López (Recorder) Salvador Gómez Pérez (Speaker)

Early in the project, a project webpage and blog was created (`http://chol.lingspace.org/`). The page is fully trilingual in Ch'ol, Spanish, and English. It contains a description of the project, as well as a link to downloadable Ch'ol resources. Student participants in the project were invited to contribute blog posts, which ranged from introductions to poetry and stories composed in Ch'ol. The blog was also used to post about events from the project, as well as other events related to the Ch'ol language in the region. The webpage now serves as a general Ch'ol resource, and was also used by a group of Cornell-based Ch'ol researchers. Authors Vázquez Álvarez and Arcos López helped set students up on the page, encouraging them to post, creating page content, translating student posts from Ch'ol to Spanish and English, as well as others' posts from English to Spanish and Ch'ol.

It is important to highlight that the students who participated in the project were in the last two or three semesters of their program. By this time, they had chosen a place to conduct research for their senior theses and had contact with people in the areas where they would conduct the research. All the interactions with people in these areas were conducted in Ch'ol. As a result, they did not face obstacles in finding people to record or in recording the narratives. Interest in the project by community members was positive and students did not have trouble finding friends and family members willing to participate.

For the completion of their majors, the students enrolled in courses that focused on their individual research projects, which included periodical fieldwork. For this reason, in particular at UAM-UNICH, the project fit in well in the academic trajectory of the students. At UAM-UNICH in Yajalón, professor Bernabé Vázquez offered service credit to the students to count towards their degree and the project offered compensation for the students' fieldwork. For the transcriptions, the participants chose materials according to the dialect of Ch'ol they knew.

One challenge faced in the project was in the consistency of the way recordings were transcribed as well as certain orthographic conventions that are not standardized across the corpus. The standard orthography for Ch'ol was agreed upon in 2010 by Ch'ol writers and bilingual teachers, based on the one proposed by the author and native speaker of Ch'ol José Díaz Peñate in Díaz Peñate (1992) and updated in INALI (2011). Nevertheless, transcriptions in this corpus are not fully consistent. Given the bulk of transcriptions, the authors decided to publish the collection with the orthographic inconsistencies, rather than waiting to clean them up and risking delaying or stymieing the publication of the collection.

Finally, we would like to highlight that AILLA was involved at every step of this project from the beginning, which facilitated greatly in organizing the collection and obtaining all the correct information and permissions for the metadata. Financial support for AILLA was written into the grant and their involvement from the beginning meant that there was additional help in the archiving process. With support from AILLA, this corpus won the SSILA Archiving Award in 2019.

3.2. Corpus 2

A second corpus, Little (2021), was created by author Little and Ch'ol-speaking research assistant Vázquez Martínez. This corpus contains recordings from two dialects, transcribed in Ch'ol and translated into Spanish. This corpus was created in part for Little's dissertation research on definiteness and word order in Ch'ol (Little 2020). Little and Vázquez Martínez recorded in San Miguel (Tumbalá dialect) and El Campanario (Tila dialect), respectively. The recordings included in this corpus were made from 2015 to 2020. This corpus was archived at AILLA as part of Little's dissertation research. Both Little and Vázquez Martínez recorded community members with whom they have known for many years. Little recorded members of the Arcos López family in San Miguel and Vázquez Martínez recorded family members and friends in her home community of El Campanario, Tila, Chiapas. Funding for costs related to travel and work transcribing and translating came from a dissertation research improvement grant from the National Science Foundation and Documenting Endangered Languages.

Each folder in the corpus at AILLA contains a recording in .wav format, an accompanying .eaf transcription file as well as metadata information. Some of the folders contain videos that were made in collaboration with community members using the recordings. The contents of the corpus are:

- 17 WAV recordings (1 hour, 37 minutes, 57 seconds)
- 18 .eaf trancriptions
- 4 .mp4 videos (14 min 19 seconds)

A sample of the audio recordings from corpus 2 are given in Table 2.

Table 2: Examples of recordings in AILLA for Little (2021)

Recording title	Location recorded	Contributors
The demons *Xi'ba*	San Miguel, Salto de Agua, Chiapas	Carol Rose Little (Recorder) Evelina Arcos López (Speaker)
The Jaguar *Bajlum*	San Miguel Salto de Agua, Chiapas	Carol Rose Little (Recorder) Rosa Arcos Pérez (Speaker)
The Mole *Baj*	El Campanario, Tila, Chiapas	Morelia Vázquez Martínez (Recorder) Augusto López (Speaker)
The Snake *Lukum*	El Campanario, Tila, Chiapas	Morelia Vázquez Martínez (Recorder) Virginia Martínez Vázquez (Speaker)

Transcriptions in ELAN were made by Vázquez Martínez and Little. All recordings have Ch'ol transcriptions and Spanish translations. Some recordings additionally contain English translations of the Ch'ol. Both authors had prior experience recording and transcribing. Little included a fully transcribed, glossed and translated narrative from this corpus in the appendix of her dissertation.

4. Applications of the corpora

In the following subsections, we detail how the corpora have been used for (i) linguistic research; (ii) creation of resources in Ch'ol; and (iii) pedagogical descriptions and materials. We also lay out how author Vázquez Martínez gave a presentation in Ch'ol on joint research on definiteness marking in Ch'ol, underscoring an instance of using the language to communicate research about the language—particularly significant for furthering the maintenance and preservation of minority languages.

4.1. Research

Both corpora have been used for linguistic research on, for example, headless relative clauses (Ander-Bois et al. 2019, Vázquez Álvarez & Coon 2021) and definiteness (Little 2020, Little & Vázquez Martínez 2018, Vázquez Martínez & Little 2020). For example, with data from the second corpus, Little & Vázquez Martínez (2018) investigated the distribution and interpretation of nouns with and without determiners in the Tumbalá dialect. They found that bare nouns could refer to unique entities and anaphoric elements. Determiners were used in anaphoric contexts, however bare nouns were still possible. Vázquez Martínez & Little (2020) expanded Little & Vázquez Martínez (2018) to include the Tila dialect of Ch'ol. Tila Ch'ol has an additional determiner not found in Tumbalá Ch'ol, *li*. Vázquez Martínez & Little (2020) found that Tila Ch'ol speakers use determiners more often, however bare nouns were still possible as definite. Some examples of bare nouns and their interpretations are given in the examples below from Tumbalá Ch'ol in (1). An anaphoric usage of the determiner *jiñ(i)* is given in (2). The italicized right aligned text refers to the narrative that the excerpt comes from.

(1) a. Context: The woman is an established protagonist.
Ta' puts'i lok'el **x'ixik**.
PFV flee away woman
'The woman fled away.' Definite *Bajlum*

b. Context: After previous mention of a fox
Ta'=bi iyajñesa **wax**.
PFV=REP chase fox
'The fox chased them, so I'm told.' Definite *Xi'ba*

c. Context: First mention of a man in the story.
Ta'=bi chämi **wiñik**.
PFV=REP die man
'A man had died, so I'm told.' Indefinite *Xi'ba*

(2) Referring back to the farmers after they were mentioned in the first line of the story.
Ya' a ta' ik'a-y-ob tyi bij a **jiñ** x-cholel-ob.
There PRT PFV spend.night-IV-PL PREP way PRT DET NC-field-PL
'The farmers spent the night on the road.' Anaphoric *Xi'ba*

This research on definiteness has contributed to a growing body of descriptive and theoretical work on how anaphoric and unique definites are expressed crosslinguistically (e.g., Deal & Nee (2017), Jenks (2018), Moroney (2021), and Schwarz (2009, 2013)).

Research based on both corpora have fed into ongoing work on definiteness and headless relative clauses, e.g., (AnderBois et al. 2019, Little et al. 2021). This research has shed light on variation across headless relative clauses, but also some robust crosslinguistic similarities with free relative constructions (see Caponigro et al. (2020) for more). Below, we will return to the importance of theoretical descriptions of linguistic properties for pedagogical materials.

Notably for the presentation at the Society of the Study of the Indigenous languages of the Americas (Vázquez Martínez & Little 2020) based on the second corpus, Vázquez Martínez gave the presentation "Dimensions of definiteness in Ch'ol: A dialectal comparison" in Ch'ol with slides in English. Vázquez Martínez and Little, with the help of author Vázquez Álvarez, translated linguistic terms such as definiteness, anaphoricity, uniqueness, generic reference, etc. into Ch'ol. A list of their translations is given in Table 3 along with the literal meaning of the Ch'ol. Some of these terms are based on those used in Jiménez Jiménez & Guzmán Gutiérrez (2013), a pedagogical grammar of Ch'ol written in Ch'ol.

Table 3: Linguistic terms in Ch'ol used in Vázquez Martínez & Little (2020)

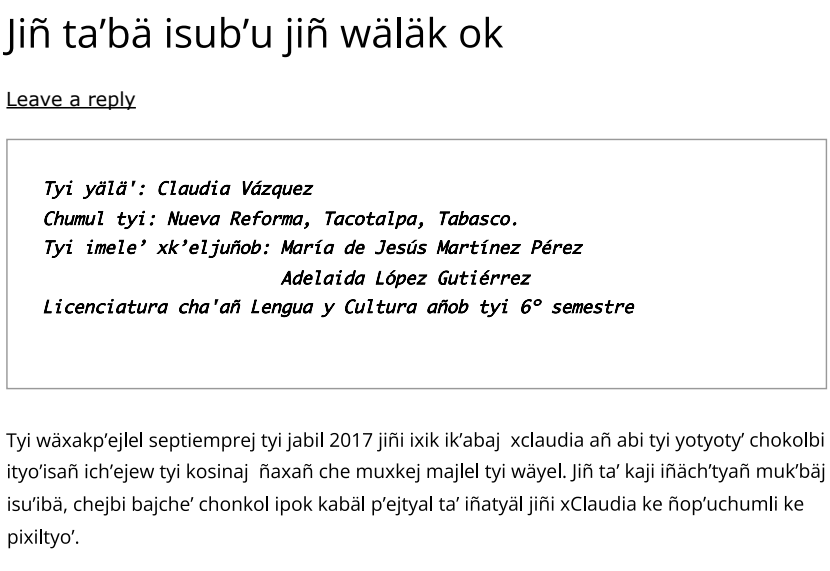
Linguistic Term	Ch'ol translation	Literal translation
definiteness	*käñbilbä ty'añ*	'known word'
definite	*käñäl*	'known'
anaphoricity	*ty'añ ñaxañ albilbä*	'aforementioned referent'
determiner	*pästyäl*	'pointer'
dialect	*yäñälbä ty'añ*	'different speech'
generic referent	*ty'añ mu'bä ichäkä ajlel*	'referent that is generally known'
unique definite	*juñsujmjach käñälbä ty'añ*	'a word with only one referent'
noun	*k'aba'äl*	'name' with the alienable suffix *-äl*
indefinite	*mach käñäl*	'unknown'
linguistics	*yomty'añ*	'collect-language'
possessor	*icha'añtyej*	'one that possesses'
transitive verb	*k'axcha'leyaj*	'across doer'
intransitive verb	*mach k'axcha'leyaj*	'does not go across'
subject	*ajcha'leyaj*	'doer'
object	*ajcha'leñtyel*	'doee'
verb	*cha'leyaj*	'thing that does'
adjective	*yilal*	'how it seems'

Presentations such as Vázquez Martínez & Little (2020) highlight the importance of conveying scientific studies on and *in* the language of investigation—especially important for the promotion of minority languages. SSILA gave the presentation Vázquez Martínez & Little (2020) a Special Recognition for incorporation of an Indigenous language in their best student presentation category.

4.2. Resources in Ch'ol

Easily accessible resources such as blog posts and short videos were created alongside the corpora. These materials serve as a way to generate more accessible content in Ch'ol. As mentioned above, the progress of the corpora were documented on the trilingual blog. The blog additionally serves as a way to generate more writing in Ch'ol on the Internet. An example of a blog post in Ch'ol detailing a story about how the Wäläk Ok, the lord of the creek, warned about an earthquake, is given in Figure 2.

Figure 2: Blog post in Ch'ol about the warning of the *Wäläk Ok* (Lord of the Creek) with metadata at the beginning `http://chol.lingspace.org/ctu/jin-taba-isubu-jin-walak-ok/`



Jiñ ta'bä isub'u jiñ wäläk ok

Leave a reply

Tyi yälä': Claudia Vázquez
Chumul tyi: Nueva Reforma, Tacotalpa, Tabasco.
Tyi imele' xk'eljuñob: María de Jesús Martínez Pérez
Adelaida López Gutiérrez
Licenciatura cha'añ Lengua y Cultura añob tyi 6° semestre

Tyi wäxakp'ejlel septiemprej tyi jabil 2017 jiñi ixik ik'abaj xclaudia añ abi tyi yotyoty' chokolbi ityo'isañ ich'ejew tyi kosinaj ñaxañ che muxkej majlel tyi wäyel. Jiñ ta' kaji iñäch'tyañ muk'bäj isu'ibä, chejbi bajche' chonkol ipok kabäl p'ejtyal ta' iñatyäl jiñi xClaudia ke ñop'uchumli ke pixiltyo'.

Along with community members of San Miguel, author Little created three videos with the recordings *Bajlum* 'The jaguar', *Empanada* 'Empanada' and *Bats'* 'The monkeys'. Community members acted in the videos to recreate the stories from these recordings. These videos are available on a vimeo page (`https://vimeo.com/user87897610`) as well as through various social media groups for Ch'ol speakers.[1] These videos have the option of subtitles in Ch'ol, Spanish and English. The videos with subtitles in Ch'ol are currently being used in bilingual education to supplement orthography classes.

4.3. Pedagogical materials

In addition to the impacts for documentation and training, careful examination of formal linguistic features leads to better materials for speakers and learners, which are, crucially, not based on the language used in educational settings (Noonan 2005, Rice 2006). Once such example comes from the work outlined above on definiteness. The influence of Spanish or English can produce inaccurate descriptions of how definiteness is marked in typologically diverse languages. As both English and Spanish have definite and indefinite articles, many pedagogical materials in Ch'ol have simply translated the definite article into a demonstrative or the anaphoric determiner *jiñi*. The indefinite article is translated into Ch'ol as the numeral 'one'. This is the case in the Spanish-Ch'ol dictionary (Aulie & Aulie 1978) and a multilingual Mayan dictionary (López K'ana et al. 2016). However, in Ch'ol there is no word directly corresponding to a definite or indefinite article (Little & Vázquez Martínez 2018, Vázquez Martínez & Little 2020). Instead, bare nouns can be definite and, as detailed in Little (2020), the order of nominals and adjunct phrases is important for the interpretation of bare nominals, as detailed in (3). When the prepositional phrase *tyi otyoty* 'in the house' intervenes between the locative predicate *ya'añ*, the bare nominal *wiñik* must be definite.

[1] One of these is the page 'Lakty'añ Ch'ol - Lengua Ch'ol - Tila', managed by the Ch'ol teacher, educator and writer, Miriam Hernández. The page can be found here: `https://www.facebook.com/LenguaChol`.

(3) a. Ya'-añ **wiñik** [tyi otyoty].
there-is man in house
'**A** man is in the house.'

b. Ya'-añ [tyi otyoty] **wiñik**.
there-is in house man
'**The**/*a man is in the house.'

The investigation of definiteness illustrates how theoretical linguistic research can be used to inform pedagogical materials.

This work is directly applicable to bilingual education in the state of Chiapas, where about a third of the population's first language is a language other than Spanish (Gobierno de Chiapas 2019). It has been shown that access to education in a child's native language promotes academic success (DeGraff 2017, Dutcher 2004, UNESCO 2016). However, in many communities in Chiapas, children are first taught in Spanish, a language many do not speak when they begin school. When children are first taught in a language they do not understand, it restrains creativity and diminishes their academic performance, ultimately leading them to devalue their native language and culture. Descriptions of grammatical properties of Ch'ol lead to increased metalinguistic awareness. Metalinguistic awareness is important for literacy and language maintenance. In Chiapas, Ch'ol is not taught formally in most schools; Spanish oftentimes is the language of instruction. Even in bilingual schools where Ch'ol is used for part of the time, there are very few resources on grammatical properties of Ch'ol. One exception is the detailed grammar, Jiménez Jiménez & Guzmán Gutiérrez (2013), a grammar of Ch'ol written in Ch'ol with didactic material and suggestions on how to teach properties of the language. The corpora and resulting research discussed here thus add to the growing body of materials in and on Ch'ol for and by Ch'ol-speakers that have clear applications to education.

5. Conclusion

We have described the creation of two Ch'ol corpora with benefits to research, resource development and pedagogical materials. Through involvement in the projects outlined above, Ch'ol students had direct ownership of the documented material, as well as the opportunity to engage with their language in different capacities. While every linguistic context is different, we believe this model can be extended to other languages. One such example of a similar project comes from Murray et al. (2020) for Cheyenne. Cheyenne is an Algonquian language spoken in Montana and Oklahoma. Ethnologue lists its status as moribund and cites 380 speakers, a statistic from 2018, making Cheyenne's linguistic context very different from Ch'ol's. Nevertheless, there are similarities in the way the project in Murray et al. (2020) was carried out to the one just described for Ch'ol. As detailed in Murray et al. (2020), student interns at the local tribal college in Lame Deer, MT worked with a database of texts and recordings in Cheyenne in order to document the distribution of demonstratives in the language. Working with mainly existing recordings, students, who were mostly from the North Cheyenne Tribe, were able to engage with their language in different capacities and gain skills in linguistic research, including a project on demonstratives in Cheyenne. The student interns also made short videos, similar to the ones made with the recordings in Ch'ol. These videos were shared on social media and were received positively by the Cheyenne community. Even though linguistic and research contexts can differ across communities, we hope to have highlighted the benefits of collaborative corpus creation and the many applications that can be made of the resulting corpora.

References

AnderBois, Scott, Miguel Oscar Chan Dzul, Jessica Coon & Juan Jesús Vázquez Álvarez. 2019. Relativas libres en ch'ol y maya yucateco y la tipología de cláusulas relativas sin núcleo. In *Proceedings of form and analysis in mayan linguistics 5*, vol. 5.

Aulie, Wilbur & Evelin Aulie. 1978. *Diccionario Ch'ol-Español, Español-Ch'ol*. Third. México: Summer Institute of Linguistics.

Caponigro, Ivano, Harold Torrence & Roberto Zavala Maldonado (eds.). 2020. *Headless relative clauses in Mesoamerican languages*. Oxford: Oxford University Press.

Deal, Amy Rose & Julia Nee. 2017. Bare nouns, number, and definiteness in Teotitlán del Valle Zapotec. In Chris Cummins, Nikolas Gisborne, Caroline Heycock, Brian Rabern, Hannah Rohde & Robert Truswell (eds.), *Proceedings of sinn und bedeutung*, vol. 21, 317–334. Edinburgh.

DeGraff, Michel. 2017. Mother-tongue books in Haiti: the power of Kreyòl in learning to read and in reading to learn. *PROSPECTS*. 1–30.

Díaz Peñate, José. 1992. *La' lakts'ijbuñ lakty'añ tyi ch'ol*. Mexico City: Gobierno del Estado de Chiapas.

Dutcher, Nadine. 2004. *Expanding educational opportunity in linguistically diverse societies*. Washington, DC: Center for Applied Linguistics.

Ethnologue. 2020. *Chol*. https://www.ethnologue.com/language/ctu.

Ethnologue. 2021. *Language Status*. https://www.ethnologue.com/about/language-status.

Gobierno de Chiapas. 2019. *Plan estatal de desarrollo Chiapas*. https://tinyurl.com/y6m7w48p. Accessed September 14, 2020.

Gómez Jiménez, Silvestre. 2015. Grupo lingüístico lakty'añ, ch'ol. In José Daniel Ochoa Nájera (ed.), *Las lenguas de chiapas*, vol. 1, 15–20. Tuxtla Gutiérrez, Chiapas: Consejo Estatal para las Culturas y las Artes de Chiapas.

INALI. 2011. *Ityoj ts'ijbuñtyel lakty'añ ch'ol [norma de escritura de la lengua ch'ol]*. México: Instituto Nacional de Lenguas Indígenas.

Jenks, Peter. 2018. Articulated definiteness without articles. *Linguistic Inquiry* 49(3). 501–536. https://doi.org/10.1162/ling_a_00280. https://doi.org/10.1162/ling_a_00280.

Jiménez Jiménez, Enrique & Jorge Guzmán Gutiérrez. 2013. *Its'ijbuñtyel ña'alty'añ ch'ol [ch'ol grammar]*. Mexico: INALI-SEP.

Little, Carol Rose. 2020. *Mutual dependencies of nominal and clausal syntax in Ch'ol*. Ithaca, NY: Cornell University Doctoral dissertation.

Little, Carol Rose. 2021. *Chol Collection of Carol Rose Little*. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: public. PID ailla:119634. Accessed May 4, 2021.

Little, Carol Rose, Scott AnderBois, Jessica Coon & Juan Jesús Vázquez Álvarez. 2021. *Super-free relatives and bare nouns in two Mayan languages: Implications for type-shifteres*. Presented at the 25th Workshop on Structure and Constituency in Languages of the Americas (WSCLA 25).

Little, Carol-Rose & Morelia Vázquez Martínez. 2018. *La distribución e interpretación de sustantivos en el ch'ol: Un estudio práctico de corpus*. Presented at Form and Analysis in Mayan Linguistics (FAMLi) 5. Antigua, Guatemala.

López K'ana, Josías, Miguel Sántiz Méndez, Bernabé Montejo López & Pablo Gómez Jiménez. 2016. *Diccionario multilingüe*. 3 (ed.). Mexico: Siglo veintiuno editores.

Moroney, Mary. 2021. Updating the typology of definiteness: evidence from bare nouns in Shan. *Glossa: a journal of general linguistics* 6(1). 1–28.

Murray, Sarah, Carol-Rose Little, Chloe Ortega, Wayne Leman, Richard Littlebear, Jessie Angel-Brien, Haley Ash-Eide & Desta Sioux Calf. 2020. *Cheyenne demonstratives: A corpus study*. Presented at the 52nd Algonquian Conference. University of Wisconsin-Madison.

Noonan, Michael. 2005. Grammar writing for a grammar-reading audience. *Studies in Language*. 351–365.

Rice, Karen. 2006. Let the language tell its story? the role of linguistic theory in writing grammars. In Felix K. Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language*, 235–268. De Gruyter Mouton.

Schwarz, Florian. 2009. *Two types of definites in natural language*. University of Massachusetts Amherst dissertation.

Schwarz, Florian. 2013. Two kinds of definites cross-linguistically. *Language and Linguistics Compass* 7(10). 534–559.

UNESCO. 2016. If you don't understand how can you learn. Policy Paper 24. Paris: UNESCO. http://unesdoc.unesco.org/images/0024/002437/243713E.pdf. Accessed September 14, 2020.

Vázquez Álvarez, Juan Jesús & Jessica Coon. 2019. *A corpus of Ch'ol narratives*. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: public. PID ailla:119634. Accessed May 4, 2021.

Vázquez Álvarez, Juan Jesús & Jessica Coon. 2021. Headless relative clauses in Ch'ol. In Ivano Caponigro, Harold Torrence & Roberto Zavala (eds.), *Headless relative clauses in languages of Mesoamerica*. New York: Oxford University Press.

Vázquez Martínez, Morelia & Carol-Rose Little. 2020. *Dimensions of Definiteness in Ch'ol: A dialectal comparison*. Presented as the Society of the Study of the Indigenous Languages of the Americas. Baton Rouge, LA.

Proceedings of the 42nd West Coast Conference on Formal Linguistics

edited by Shweta Akolkar,
Amber Galvano, Akil Ismael,
Kang Franco Liu, and Line Mikkelsen

Cascadilla Proceedings Project Somerville, MA 2025

ISBN 978-1-57473-484-3 hardback

Ordering information

Orders for the printed edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Little, Carol Rose, Juan Jesús Vázquez Álvarez, Jessica Coon, Nicolás Arcos López, and Morelia Vázquez Martínez. 2025. Collaborative Corpus Creation: A Ch'ol Case Study. In *Proceedings of the 42nd West Coast Conference on Formal Linguistics*, ed. Shweta Akolkar et al., 478-487. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3853.