# Collaborative corpus creation: A Ch'ol case study

*Carol Rose Little[1], Juan Jesús Vázquez Álvarez[2], Jessica Coon[1], Nicolás Arcos López[3], and Morelia Vázquez Martínez*

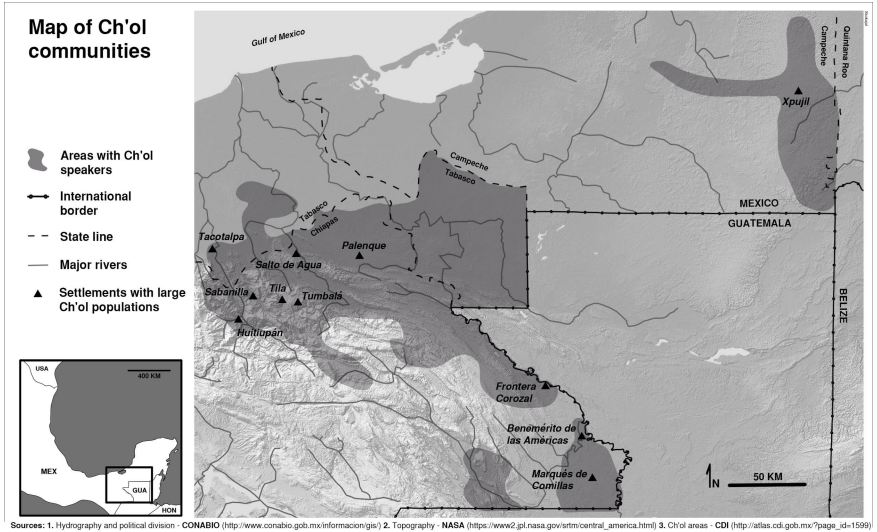
[1]McGill University, [2]CIMSUR-UNAM, [3]UIET

WCCFL 39
April 10, 2021
The University of Arizona

# Today

- **Goal:** describe two collaborative linguistic research and documentation projects which created two corpora of Ch’ol narratives, available through the Archive of Indigenous Languages of Latin America (AILLA; ailla.utexas.org).

- Ch’ol is a Mayan language spoken in Southern Mexico by about 252,000 people.




# Today

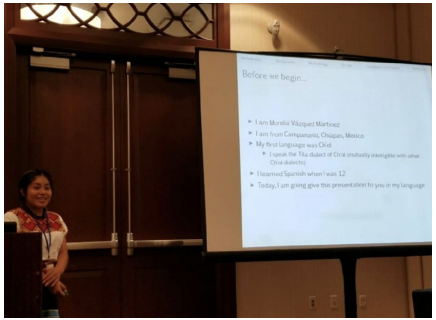
- The two projects served twin goals of:
  - facilitating linguistic research and creating documentation materials on Ch’ol
  - increasing language awareness and building capacity among Ch’ol-speaking students, who were involved in all stages of the project.
- We discuss how this “crowd-sourcing” approach to linguistic corpus creation has the potential to benefit both language communities and researchers.
- First we outline the process of workshops which resulted in one of the corpora, then we highlight how materials from both corpora have been used in linguistic research, and discuss benefits to communities.

# Who we are

- The presenters of this paper represent the different roles of participants:
    - Native-speaker linguists working in universities in southern Mexico;
    - linguists in the US and Canada;
    - a Ch’ol-speaking student who participated in the workshops and both corpus projects.

Nicolás Arcos López

Carol Rose Little




Morelia Vázquez Martínez

Juan Jesús Vázquez Álvarez

Jessica Coon

# Juan Jesús Vázquez Álvarez

- Speaker of Ch’ol (Tila dialect)
- Associate researcher at CIMSUR-UNAM, San Cristóbal de las Casas, Chiapas, Mexico
- Research with Ch’ol language and culture
- My first description about the verbal morphology of Ch’ol (2002) opened a deeper study on aspects of Ch’ol grammar under a descriptive and theoretical point of view.
- A grammar of the Ch’ol language published as Vázquez Álvarez (2011)

# National Geographic Project

- At the end of 2017, with Jessica Coon, we discussed the best way to collect data from the Ch’ol municipalities.
- We decided to include students from two universities, completing a degree in language and culture:
  - **UIET**: Universidad Intercultural del Estado de Tabasco
  - **UAMY**: Unidad Académica Multidisciplinaria de Yajalón, Universidad Intercultural de Chiapas

# Workshops

- We established communication with one profesor from each university, who had linguistic training.
- We included workshops on aspects of documentation and linguistic diversity.

# Recording and transcribing

- The students who participated in the workshops headed to their home communities to record friends and family members on the weekends.
- They received financial support for their participation in the project and compensation for their travel costs in the recording process; some received service credit at their universities.
- Our corpus contains material from a number of different Ch'ol communities, spanning the major dialect regions.
- Students who chose to continue from both the Tabasco and Chiapas groups convened in San Cristobal de las Casas (hosted at CIESAS-Sureste) for a two-day ELAN transcription, which included training in Ch'ol orthography and writing.

# Teaching the structure of Ch’ol in the classroom

- The corpus is useful for communities, for non-speakers, for students, and researchers.
- It provides material for teaching the language to non-speakers.
- The corpus is being used to teach aspects of the Ch’ol grammar to teachers in bilingual schools.
- The corpus helps to contextualize and educate speakers about dialectal variation.
- Makes students understand the value of Ch’ol and understand it on a more meta level.

# Nicolás Arcos López

- Professor of languages and cultures at the Intercultural University of Tabasco (UIET)
- Speaker of Ch’ol (Tumbalá dialect)


# Workshop on documentation at UIET

- Occurred on February 9, 2018 at UIET and covered the following:
  - Introduction on endangered languages
  - The importance of documenting endangered languages for communities
  - Documentation and revitalization of languages
  - How to record audio and videos
  - The context of Ch’ol and our project with National Geographic


# Workshop on documentation at UIET

- Professors and students from the department of languages and cultures at UIET participated in the workshops
- They learned about the technical aspects of recording (i.e., how to reduce background noise, collecting metadata)
- Tutorial on ELAN for the transcriptions
- 5 students, all Ch'ol speakers, from UIET went to their communities (in Tila, Chiapas and Tacotalpa, Tabasco) to record and then transcribe Ch'ol narratives
- These narratives are now archived at AILLA


# Morelia Vázquez Martínez

- I am from El Campanario, Mexico
- My first language is Ch’ol
- I began working in linguistics in 2015
- I will talk about my involvement in the creation of a second corpus and my work investigating definiteness and the distribution and interpretation of nouns with and without determiners

# Corpus creation

- In 2018, I began recording narratives in Ch'ol in El Campanario
- After transcribing and translating narratives from El Campanario (Tila dialect) and San Miguel (Tumbalá dialect), Carol-Rose and I investigated how each dialect marks definiteness
  - We coded for instances of nouns with and without determiners and demonstratives
- In 2020, we presented our findings at SSILA: we found that Tila speakers use determiners more often in definite contexts (Vázquez Martínez & Little 2020)
- Bare nouns occur as definite more often in the Tumbalá dialect

| Story | Sentence | Noun | D/I | Det | S/O? | Order | A/U | Pred | |
|---|---|---|---|---|---|---|---|---|---|
| kajpe' | ¿ichoch mi ak'äñ li rok ta' che'iñi? | li rok | d | li | o | vo | a | t | previous sentence: mi kch'äme'maj kchi yik'oty krok (virginia) |
| kajpe' | ¿y jiñku jiñi kajpe'i säk'bil o ch'ajach mi a wa' tyikisañla? | jiñi kajpe'i | d | jiñi | s | sp | a | t | conversation about coffee |
| kajpe' | wa'li che mukbä la' juch' jiñ kajpe' mach sumuk bajche jiñi | jiñ kajpe' | d | jiñi | o | vo | a | t | conversation about coffee |
| lukum | chi'iñ kaña kwuty tyi ñumi tyi maja' li lukum | li lukum | d | li | s | vs | a | i | talking about the snake that is known to both speakers |
| lukum | mismo jiñäch ta' yubi ajkoralillo | aj coralillo | d | no | s | vs | a | i | established the snake is a coralillo |
| lukum | ya' meku kuk'ux jolo li lukumi, pero tyoj letse kx'ijal, | li lukumi | d | li | s | vs | a | a | said after añ lukum |

Unique definites

Table 5: Tumbala unique definites

| | |
|---|---|
| Bare | 14 |
| Total | 14 |

Table 6: Tila unique definites

| | | | |
|---|---|---|---|
| Bare nouns | | | 5 |
| With a determiner | *li* | 12 | |
| | *jiñi* | 2 | |
| | | | 14 |
| | | Total | 19 |

Screenshots from Vázquez Martínez & Little (2020)

# The importance of this work

- Our corpus is being archived at AILLA
- Can be used for speakers, learners, scholars
- Written record of the language
- Important to do dialectal comparisons so that speakers learn to appreciate dialectal differences rather than judge

# Carol Rose Little

- Currently a postdoctoral fellow at McGill University, previously graduate student at Cornell University
- Work with Ch’ol since 2015
- I will share outcomes from my dissertation work and the co-creation of another multidialectal corpus at AILLA and how it has been used for research on definiteness and extensions to pedagogical materials.
- I will conclude with an example of how this project has been used to create media in Ch’ol.

# Second corpus

- The second corpus was created for my dissertation.
- The creation of this corpus led to joint work on definiteness, with implications for theories of bare nouns and definiteness marking (Jenks 2018, Moroney To Appear).

Anaphoric context: After previous mention of *x-k'aläl* 'girl' twice before:

*K'uñtya k'uñtya mi y-ust-es-äñ-tyel i-tyaty i-ña' jiñi*
slow slow IPFV A3-convince-CAU-DTV-PSV-NML A3-father A3-mother DET
*x-k'aläl.*
NC-girl

'Slowly the parents of the girl become convinced.' (Cuentos Cultura Chol: 14)

Little (2020)

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

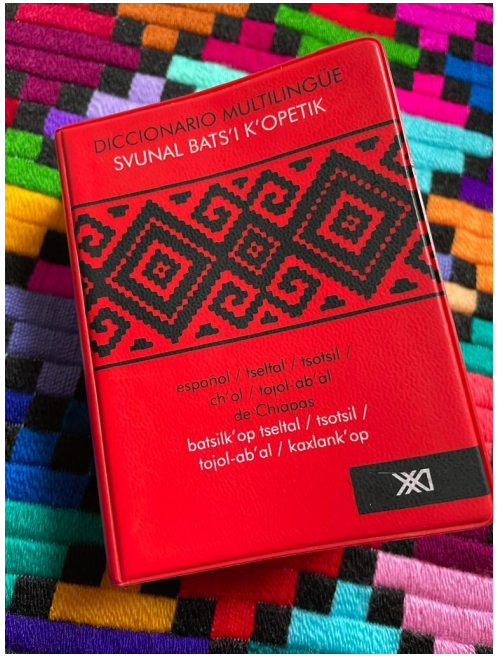
Excerpts from the narrative Bajlum from corpus 2

Table 7: Tumbalá anaphoric definites

| | | | |
|---|---|---|---|
| Bare nouns | | | 41 |
| With a determiner | *aj(iñi)/jiñi/je'* | 70 | |
| | *li/ili* | 3 | |
| | | | 73 |
| | | Total | 114 |

Vázquez Martínez & Little (2020)

# Implications for pedagogical materials




**ejote** *s* [tsel] pajk'en [tsots]
pak'ayom [ch'] patybu'ul
[toj] yaxal chenek'

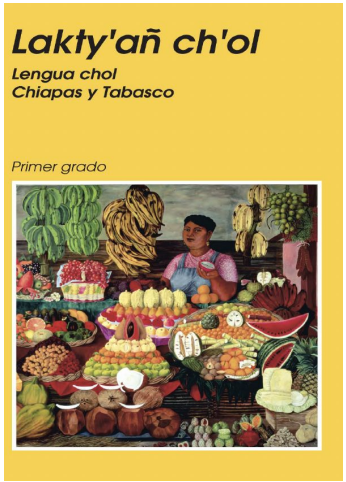
**el** *art* [tsel] te [tsots] li [ch']
jiñi [toj] ja, ye'n

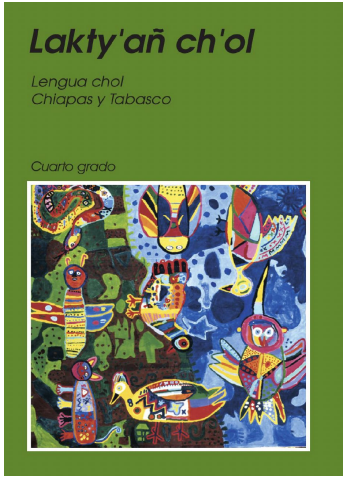
**el** *pron* [tsel] ja' [tsots] ja'
[ch'] jiñi [toj] ma', ye'na

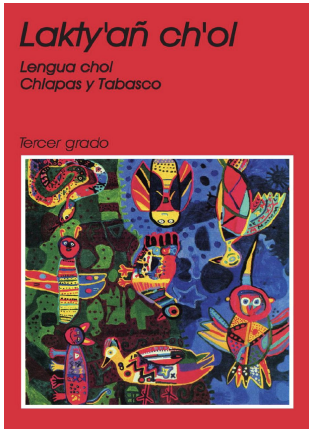
**elaborar** *vt* [tsel] pasel
[tsots] pasel [ch'] mel,

# Implications for pedagogical materials

- For example, the influence of Spanish or English can produce inaccurate descriptions of how definiteness is marked in typologically diverse languages.
- As both these languages have definite and indefinite articles, many pedagogical materials in indigenous languages simply translate the definite article into a demonstrative, and the indefinite article as the numeral ‘one’.



# Conclusions

- 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.
- In addition to the impacts for documentation and training, careful examination of formal linguistic features, such as definiteness, leads to better materials for speakers and learners, which are, crucially, not based on the language used in educational settings (Spanish, in this case).
- Capacity building amongst Ch'ol speakers
    - Model can be extended to other languages, see also a similar project with Cheyenne in Murray et al. 2020

# Conclusions

- Materials can be used to create other media, we will conclude with one such example.


https://vimeo.com/282342925


¡Wokolix la’wälä!



We would like to thank students and professors at the Universidad Intercultural del Estado de Tabasco (UIET) and La Unidad Académica Multidisciplinaria Yajalón (UAM) de la Universidad Intercultural de Chiapas (UNICH) 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.

We are also humbled to have received a SSILA archiving award for the National Geographic Project and a SSILA best student presentation award for Special Recognition for the incorporation of an indigenous language for Vázquez Martínez & Little (2020).

NATIONAL GEOGRAPHIC

NSF

Engaged Cornell

# References


Coon, Jessica and Juan Jesús Vázquez Álvarez. Chol Collection of Juan Jesús Vázquez Álvarez andJessica coon. The Archive of the Indigenous Languages of Latin America. https://ailla.utexas.org/islandora/object/ailla:261383

Jenks, Peter. (2018). Articulated definiteness without articles. *Linguistic Inquiry,* 49(3), 501-536.

Little, Carol-Rose, and 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.

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

Moroney, Mary. Accepted. Updating the typology of definiteness: Evidence from bare nouns in Shan. *Glossa.*

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

Vázquez Álvarez, Juan Jesús. 2011. A grammar of Chol, a Mayan language. Austin, TX: University of Texas Austin Doctoral dissertation.

Vázquez Martínez, Morelia, and 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.