Corpora in the time of Cholera: The pandemic's effects on language documentation

Irina Burukina & Polina Pleshak*

Abstract. This article presents a documentation project aimed at creating an open-access corpus of Patzún Kaqchikel, an endangered Mayan language spoken in Chimaltenango, Guatemala. Started in 2019 in collaboration with the Patzún Women's Cooperative Aj Su'm, the project originally sought to produce a trilingual (Kaqchikel–Spanish–English) book of recipes and oral histories. The COVID-19 pandemic and subsequent travel restrictions forced us to pivot and present the collected recordings as a dual-format online archive: (i) a research-oriented corpus with time-aligned transcriptions, translations, and morphological glossing, and (ii) a community-oriented web page featuring audio recordings, minimally-edited Kaqchikel transcripts, and Spanish translations. The collection includes over eight hours of semi-structured interviews recorded with 17 speakers of varied social and occupational backgrounds. Although the pandemic shifted the workflow toward a more academia-centered model, the project demonstrates that collaborating with the community can simultaneously satisfy local and scholarly needs and enhance the value of language documentation for both speakers and linguists.

Keywords. Language documentation, corpus, fieldwork, Kaqchikel, Mayan, COVID-19

1. Introduction. This paper focuses on the challenge of creating a corpus of an endangered language that would be of equal value to local communities and to linguists. We present a project documenting Patzún Kaqchikel, a variety of Kaqchikel (Mayan) spoken in Patzún, Chimaltenango department, Guatemala, and discuss the challenges it faced during the COVID-19 pandemic and the methodological adaptations that followed. Started in 2019 in collaboration with the Patzún Women's Cooperative Aj Su'm, the project originally aimed at producing a trilingual (Kaqchikel–Spanish–English) community-oriented book of recipes and oral histories. Because of the travel restrictions introduced in 2020 and the subsequent shift to remote work, the project was restructured to create an open-access corpus of spoken Kaqchikel instead.

The resulting collection contains over eight hours of narratives in Kaqchikel recorded from 17 speakers with diverse social and occupational backgrounds. It is available in two formats: (i) a bilingual community-oriented website with audio recordings, minimally edited transcripts in

* This project would not have been possible without the help of Filiberto Patal Majzul, who carried out the transcription and translation of the recordings. We are also deeply grateful to Maria Polinsky for her constant support at every stage of the project. We thank Pedro Mateo Pedro for facilitating our research in Guatemala and Ana López and members of the Women's Cooperative Aj Su'm for arranging interviews with Kaqchikel speakers. We are indebted to the members of the Patzún community who generously shared their knowledge and their life experiences with us. We also thank the anonymous reviewers and the audience at the 2026 annual meeting of the Society for the Study of the Indigenous Languages of the Americas (SSILA) for their insightful comments and questions.
The research was supported by Jacobs Research Funds (two grants awarded in 2019 and 2020) and Endangered Language Funds (grant awarded in 2020). Additional support was provided by crowdfunding contributions via GoFundMe.com from individual donors; we would like to thank all the contributors for their help.
Authors: Irina Burukina, University of Florida (irinaburukina@ufl.edu) & Polina Pleshak, ELTE Hungarian Research Centre for Linguistics (polinapleshak@gmail.com).

Kaqchikel, and Spanish translations, and (ii) an annotated linguistic corpus with time-aligned transcriptions, translations in Spanish and English, and morphological glossing.

While many language-documentation projects have recently shifted from primarily academic goals toward stronger community engagement, our project followed the opposite trajectory. Conceived as a community-driven initiative, it was forced by pandemic restrictions to move toward a more linguist-centered workflow. The project demonstrates that community-oriented design can enrich academic outcomes: even under constrained conditions, collaboration produces materials that bring value both to the local speaker community and linguistic research.

2. Language background. Kaqchikel is a Mayan language from the K'ichean-Mamean (Eastern) branch (Kaufman 1974). According to the 2019 census, it is spoken by approximately 410,000 people in Guatemala, most of whom are bilingual in Kaqchikel and Spanish (Richards 2003); it has been classified as threatened (Eberhard et al. 2022) or vulnerable (Moseley 2010). The present project focuses on documenting the variety of Kaqchikel spoken in Patzún (Chimaltenango department, Guatemala), where our team conducted fieldwork between 2017-2019 under the auspices of the Guatemala Field Station of the University of Maryland (more on that in section 3).

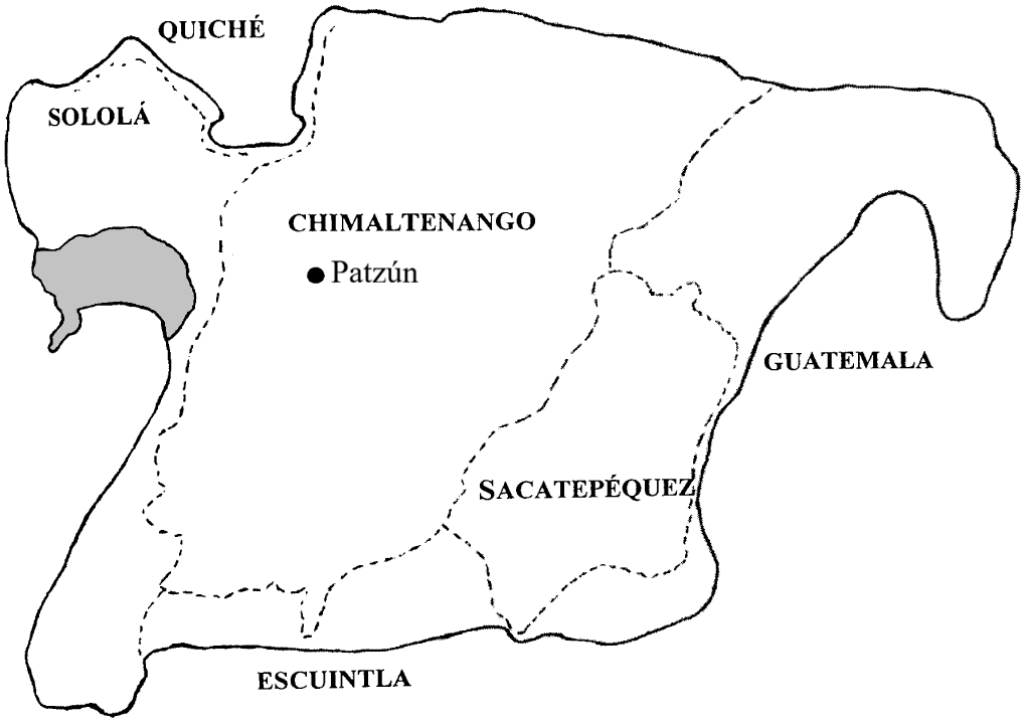

Figure 1. Regions in Guatemala where Kaqchikel is spoken[1]

For a discussion of the unique linguistic characteristics of Patzún Kaqchikel and how it differs from other regional varieties, we refer the reader to Patal Majzul et al. (2000). From a sociolinguistic perspective, language use in Patzún varies significantly across generations, even though the 2003 Ley de Idiomas Nacionales (Decree 19) recognizes and promotes the use of Guatemalan indigenous languages, including Kaqchikel. Older speakers typically use Kaqchikel

[1] Based on an image from Patal Majzul et al. (2000: 167). The area shaded grey is Lake Atitlán. The doted lines indicate the boundaries between different departments (regions). Guatemala here refers to Guatemala Department, where the capital of the country is located.

in everyday communication and often feel more comfortable speaking Kaqchikel than Spanish. In our experience, their spontaneous conversations in Kaqchikel were considerably richer than the communication with the researchers in Spanish. Middle-aged speakers tend to be balanced bilinguals, using both languages with similar ease. Younger speakers, by contrast, often prefer Spanish in everyday life, and many children are receptive speakers of Kaqchikel but do not speak it actively; see also, e.g., Heinze (2004) on bilingualism among Kaqchikel children, as well as Romero (2017) and references therein on the sociolinguistics of Mayan languages, highlighting the dominance of Spanish.

This process of language shift can be attributed to the perceived socioeconomic value of proficiency in Spanish, which is widely considered necessary for employment opportunities. Consequently, many parents reduce the use of Kaqchikel with their children, weakening intergenerational transmission of the language.

At the same time, local community members report that the presence of linguists studying Kaqchikel in Patzún has positively influenced the general attitude toward the language, increased its perceived value, and boosted greater use of Kaqchikel among younger children (Stephen Alajajian, p.c.). While the long-term effects remain to be studied, this shift is encouraging as it reflects the potential impact of collaborative documentation projects.

An additional important characteristic of Patzún Kaqchikel is the widespread use of code-switching between Kaqchikel and Spanish, including the use of Spanish nonce borrowings in otherwise Kaqchikel discourse (Romero 2015 and references therein). Speakers often demonstrate high metalinguistic awareness of this phenomenon; in interviews and elicitation sessions, many participants deliberately use Spanish lexical items in contexts where they would normally appear in natural speech, even when aware of the corresponding Kaqchikel equivalents.

3. Project design. The project began in summer 2019 and aimed at creating a collection of Kaqchikel narratives covering topics relevant and engaging to the community. The project team consisted of two PIs, Irina Burukina and Polina Pleshak, who by that time had several years of experience conducting fieldwork on Mayan and non-Mayan (Uralic, Caucasian) languages.[2] The fieldwork infrastructure in Patzún was provided by the Guatemala Field Station, supported by the University of Maryland and supervised by Drs. Pedro Mateo Pedro, Maria Polinsky, and Omer Preminger (Polinsky 2019). Funding for the project was generously provided by Jacobs Research Funds (2019, 2020) and Endangered Language Funds (2020), through three grants to the PIs; additional support was provided by crowdfunding contributions via GoFundMe.com from individual donors, who had no role in the project's design and data collection (see also the first footnote).

The project's ultimate goal was to use the collected data to prepare and publish a trilingual (Kaqchikel–Spanish–English) book of recipes and oral histories in collaboration with the Patzún Women's Cooperative Aj Su'm; a publisher in the US—TBR Books (Calec)— expressed their interest in the project.

From the beginning, we used an immersive approach to fieldwork, promoted by the Guatemala Field Station (section 3.2), and heavily relied on working together with the community, particularly with members of the Cooperative. Their help with topic selection, recruitment of

[2] In 2019, Irina Burukina obtained a PhD degree from the Eötvös Loránd University in Budapest, Hungary, and joined its Department of English Linguistics as an assistant professor. The same year, Polina Pleshak joined the Department of Linguistics at the University of Maryland as a graduate student.

interviewees, and decisions on the format of the materials was invaluable and ensured local relevance.

In summer 2019, we spent three weeks in Patzún conducting semi-structured interviews and recorded approximately ten hours of spoken Kaqchikel from seventeen speakers representing a range of occupational and educational backgrounds. Rather than attempting to cover all possible topics, we deliberately avoided genres that our consultants found challenging and that were less representative of contemporary language use, such as traditional songs and fairy tales. Instead, we prioritized the following genres: recipes of common meals, current local customs, modern life in the community, and oral history of the Guatemalan Civil War of the 1990s. Section 3.1 discusses the topics in more detail, while section 3.2 focuses on the project's data-collection methodology.

3.1. CHOICE OF TOPICS. To collect ecologically valid language data that reflected contemporary speech patterns, we relied on the guidance of community members when selecting the topics. First, we asked members of the Cooperative to list aspects of the Kaqchikel culture and everyday life in Patzún that they considered important. Then, they suggested speakers of Kaqchikel whose experience was relevant and whom we should interview. This approach ensured both community relevance and personal engagement, as the consultants discussed issues that they knew well and genuinely cared about. We further relied on word-of-mouth recommendations: some interviewees introduced us to other prospective speakers, establishing a sense of inherited familiarity.

Because of our collaboration with the Women's Cooperative, the first interviews focused primarily on women's experiences—traditional crafts, childbirth, motherhood. We then expanded the project to include men's perspectives, interviewing Mayan and Catholic religious leaders, which provided a comprehensive picture of social life in Patzún. Many interviews complemented each other in terms of viewpoints. For example, after recording a narrative about traditional medicine, we conducted a contrasting interview with a local hospital nurse. Similarly, after observing a traditional Mayan ceremony and interviewing a Mayan ritual specialist, we met with a Catholic priest.

Most interviews organized with the help of members of the Cooperative present extended narratives produced in response to questions posed by the PIs in either Kaqchikel or Spanish. We did not use a detailed questionnaire; instead, we clarified details and followed up on issues that emerged naturally during the conversation. Our questions are excluded from the publicly available corpus and from the materials intended for the community.

The collection also includes informal conversations recorded within the host families. As we engaged in daily activities—joining women to cook or accompanying them to the market—we expressed our interest and asked questions. When a speaker ~~showed~~ willingness to share, we encouraged them to discuss the activity in more detail for recording the responses when they had time.

The resulting collection covers the following topics:

- recipes,
- traditional embroidery,
- traditional music, in particular marimba (a musical instrument popular in Guatemala and throughout Central America),
- traditional medicine,
- motherhood, child-rearing,
- women-owned local small businesses (a shop),
- Women's Cooperative Aj Su'm,

- Mayan center (Centro de Estudio Maya),
- Kaqchikel evening school,
- traditional Mayan religion,
- Catholicism in Patzún,
- the Guatemalan Civil War (1960-1996),
- Iximche', the capital of Late Postclassic Kaqchikel Maya kingdom, located in Chimaltenango Department.

3.2. TEXT COLLECTION METHODOLOGY. During fieldwork, the PIs lived with local families and participated in daily activities such as cooking, crafts, household chores, and community festivals, which gave us a valuable insight into local practices and interests. Additionally, each summer between 2017 and 2019 we took two-week immersion classes in Kaqchikel supported by the Guatemala Field Station, Maya Health Alliance Wuqu' Kawoq, and the Cooperative. Together, these factors allowed us to build trusting relationships with the community, especially since none of the team members is a native speaker of a Mayan language or was born or lives in Guatemala. They further enabled more natural interactions, prompting the interviewees to share their experiences more freely.

In long-term projects, researchers may eventually acquire a high level of proficiency in the target language. In our case, this was not feasible: we had started working on Kaqchikel only a short time before the launch of this project; we could travel to Guatemala once a year, and each visit lasted no more than a month. However, intensive courses in Kaqchikel allowed us to follow the general flow of narratives and react appropriately, using both Kaqchikel and Spanish to ask questions.

A key methodological decision was to view the local community as the primary audience for the collected narratives. This strategy mitigated our limited proficiency in Kaqchikel. When native-speaker consultants realize that the immediate listener (the researcher) does not fully understand their language, they often adjust their speech to make it more accessible to an "outsider", thereby reducing narrative richness. In contrast, envisioning a fully proficient community member as the intended listener, the interviewees spoke with greater ease, producing richer narrative styles than are typical in elicited sessions.

4. Challenges faced in the project.

4.1. THE ORIGINAL TIMELINE. Prior to the COVID-19 pandemic, the project's timeline comprised three stages:

- Summer 2019: fieldwork in Patzún to strengthen relationships with the community and record the first interviews (Stage 1)
- Summer 2020: follow-up fieldwork to transcribe and translate the recordings and gather additional texts and visual materials (Stage 2)
- End of 2021: preparation of a manuscript of a trilingual book of recipes and oral histories in Kaqchikel, Spanish, and English, based on the collected materials (Stage 3).

Stage 2—transcription and translation—was to be carried out with native-speaker consultants and to follow a three-step procedure established in the Lomonosov Moscow State University's fieldwork tradition (Kashkin 2020: 5).

Step I. An audio recording is played fragment by fragment, and the consultant repeats each fragment slowly until the linguist can produce an accurate transcription; the consultant then provides a Spanish translation.

Step II. The linguist independently re-listens to the recordings, verifies that each transcribed segment accurately reflects the audio signal, and identifies gaps or inaccuracies in the transcription, translation, and potential glossing.

Step III. The linguist and the consultant jointly review the problematic fragments and resolve any remaining questions.

Stage 3—preparation of the manuscript—was to involve editing the Kaqchikel texts and standardizing the orthography, confirming the Spanish translations, and adding English translations. We expected to revise the original texts by removing hesitation markers and repeated fragments while keeping most of the original content. The edited texts would then be reviewed by a native Kaqchikel speaker to confirm that the editing had not disrupted the natural flow of the language. The Spanish translations would be checked by a native Spanish speaker, and the English translations would be produced by the PIs and proofread by a native English speaker. The completed manuscript would have been submitted to the publisher in late 2021 or early 2022.

4.2. UNEXPECTED CHALLENGES. The COVID-19 pandemic introduced a number of unforeseen challenges to the original plans. The most significant was our inability to return to Patzún in the summer of 2020 to conduct in-person transcription and annotation sessions because of pandemic-related lockdowns and travel restrictions. Consequently, we could not conduct additional interviews, collect visual materials for the book, or clarify certain details in the recipe narratives.

These constraints forced us to reconsider the project's goals and timeline. In particular, we asked:

- Should we collect additional texts remotely or focus on the existing recordings?
- Should we aim to obtain high-quality visual materials for the book?
- How could we carry out transcription and translation without in-person meetings?

Given that we already had a substantial amount of recorded material, we decided to focus on processing the existing audio recordings rather than attempting to expand the dataset.

At the same time, the collection contained gaps that made preparing a book manuscript difficult. Kaqchikel culinary discourse describes food preparation differently from the expectations associated with written recipes in Western publishing traditions. In particular, the recipes in our corpus lack precise lists of ingredients and quantities, which are often considered an indispensable part of a recipe in the Western world. In Kaqchikel meal descriptions, the focus is on the actions as well as on the context in which the dish is cooked and its possible variations. Despite our best efforts to elicit additional details during the interviews, the resulting narratives are closer to cooking descriptions rather than to step-by-step instructions.

During the 2020 fieldwork season, we intended to prepare the described meals together with members of the Cooperative and our host families, clarifying the quantities through hands-on cooking and taking photographs for later use in the book. When travel became impossible, we had to abandon these plans. The immediate priority shifted to processing the existing recordings, and we decided not to pursue detailed recipe reconstruction or additional visual material. Instead,

we chose to preserve the narratives as they were naturally produced, thus reflecting the community's vision.

Transcribing and translating the audio recordings therefore became our main task. We considered two strategies. First, we could organize synchronous online transcription sessions with native-speaker consultants to replicate the traditional workflow as closely as possible. Second, we could employ a Kaqchikel-speaking language specialist to carry out transcription and translation independently in an asynchronous format.

The first option required stable, high-quality internet access and appropriate audio equipment on both sides. Based on our communication with community members during the pandemic, we concluded that such conditions could not be reliably guaranteed. In addition, community members understandably prioritized their work and families and could not commit to regular online meetings.

The second option—asynchronous transcription by a trained native speaker—was more feasible. We were fortunate to work remotely with Filiberto Patal Majzul, a native speaker of Kaqchikel, a trained linguist, and (co-)author of several important publications on Kaqchikel grammar and lexicon, including Patal Majzul et al. (2000) and Patal Majzul (2007). His extensive linguistic background and familiarity with transcription and annotation tools made this remote workflow possible.

4.3. EFFECTS OF ASYNCHRONOUS REMOTE WORK. Delegating the transcription process to a single native-speaker assistant inevitably slowed the overall work process. According to the original plan, the PIs would have worked in parallel with two native speakers, doubling transcription speed. We also planned to recruit multiple consultants to allow longer daily work sessions. Transcription is cognitively demanding for speakers and requires regular breaks, but working several hours a day with different people could have completed this stage of the project within three to four weeks of intensive fieldwork. With only one native-speaker linguist working remotely, this part of the project had to be reconsidered.

Remote communication introduced additional delays. When working on site, transcription sessions are typically organized in a fixed daily schedule. In contrast, asynchronous communication relies on deadlines that may be weeks apart, which can slow progress. Our initial estimate was to allow about two months for transcribing and translating one hour of audio, but it often took much longer due to unforeseen changes in the personal lives of the members of the team. Fortunately, we did not experience problems with international wire transfers and were able to compensate Filiberto Patal Majzul promptly; the rate had been discussed with a supervisor at the Guatemala Field Station and agreed upon with the language specialist.

In addition to the general lack of manpower and higher time costs of asynchronous work, the project's timeline was affected by inevitable delays in responding to clarification questions. In an in-person transcription/translation session, most if not all questions can be resolved immediately. Online communication is more time-consuming; questions sent by email may be addressed later and sometimes only partially.

As a result of the changes and challenges discussed above, the planned timeline changed substantially. The book publication was deferred and our focus gradually shifted toward the most achievable outcome under the present circumstances: developing an open-access annotated corpus of Patzún Kaqchikel narratives and creating a separate web page to make the recordings available to the community. With the rapid introduction of new online tools and resources during the pandemic, this direction aligned well with broader trends in linguistic research and digital archiving.

5. Adjustments.

5.1. REVISED OBJECTIVES. Since we were unable to collect additional information about the recipes or record new interviews, we decided instead to focus our efforts on maximizing the value of the existing recordings and preserving their natural character.

The recorded interviews about everyday life in Patzún are linguistically rich and include a wide variety of tense-aspect combinations, finite and non-finite constructions, and discourse strategies such as topicalization and focalization. Although the collected recipes may lack the precise details required for publication as conventional cooking instructions, they serve as valuable illustrations of natural language use. These texts demonstrate narrative organization and code-switching practices and provide examples of various grammatical constructions, such as purpose clauses, imperatives, and coordination.

The recordings are fully transcribed and translated into Spanish, with hesitation pauses and other features common to spontaneous speech marked. We preserved the orthography suggested by Filiberto Patal Majzul. At the same time, to make the corpus easier to search, we are currently adding a second transcription tier using the standard orthography presented in the dictionary by Patal Majzul (2007); this tier includes only items where we noted discrepancies, e.g., *vave'* vs. *wawe'* 'here'. We expect the entire corpus to be available online within the calendar year 2026.

The collection is distributed in two formats aimed at the two intended audiences. For researchers, the recordings are available as time-aligned ELAN (.eaf) files accompanied by translations; additionally, a morphologically annotated subcorpus of approximately 3000 tokens with added English translations is available in FieldWorks (.flex) format. For the community, we have been developing a dedicated bilingual web page containing the recordings, lightly edited Kaqchikel transcripts (free from the most obvious repetitions and marked pauses), and Spanish translations.

This dual-format approach allows the materials to serve both academic and community needs while respecting the different expectations of these audiences.

5.2. REVISED STRATEGIES. When preparing the audio recordings for the corpus, we first edited them by removing meta-conversations, the interviewers' questions, and outside noise, and by isolating and preserving the relevant narrative segments in Kaqchikel. The materials were then divided into manageable portions of approximately one hour of audio each and sent one at a time to Filiberto Patal Majzul for transcription and translation into Spanish. As mentioned above, we expected each portion to be completed within about one to two months. Upon receiving the transcription, we verified its completeness and transferred payment before sending the next hour of audio.

Initially, transcriptions were produced in Microsoft Word, with each line accompanied by start- and end-timestamps. These files were then manually transferred into ELAN by the PIs. While this workflow allowed convenient commenting and discussion of unclear segments, it proved unnecessarily time-consuming. Subsequently, transcription was carried out directly in ELAN, allowing the language specialist to annotate audio segments more efficiently and eliminating the need for manual alignment. It also did not preclude commenting: we (the PIs) created a separate list of questions, in which we noted the comment, file name, and time boundaries of each relevant segment, and later shared this list with Filiberto Patal Majzul for feedback. At the same time, the asynchronous transcription process led to some disagreement regarding segmentation of the audio recordings. The resulting transcriptions do not consistently reflect clause boundaries or pauses in the speech signal. We acknowledge that this issue needs to be addressed

in the future; however, correcting it in the entire corpus will require substantial additional time and manpower that are currently limited for our team.

Each transcript was reviewed by at least one PI to ensure that it fully corresponded to the recording and to eliminate discrepancies between the transcription and the audio. We further added annotation tiers containing dialectal and cultural commentary, following the convention used in the online open-access Kaqchikel corpus by Bennett & Henderson (2022), which serves as the model for our collection.

As mentioned above, we further had to decide on standardization of the orthography in the transcription. We contend that dialectal and phonetic variation should be preserved rather than normalized, and we created a separate tier in the .eaf files for the standard forms whenever there is a mismatch. Another issue worth addressing was vowel reduction. Initially, we wanted to preserve transcriptions with omitted vowels as prepared by Filiberto Patal Majzul. However, upon reviewing the transcriptions and listening carefully to the audio signal, we noticed several likely mismatches between the audio and the text, though our level of proficiency in Kaqchikel makes our judgments less reliable. The frequent omission of vowels complicates corpus searches, and we are currently contemplating either indicating vowels in the original transcription in contexts where they are clearly audible or adding the standard forms in the "standardized" tier for all instances of reduced realization.

The ELAN files, specifically the Kaqchikel and Spanish tiers, were further exported to plain-text format and lightly edited for the community website. The editing included removing repeated fragments, hesitation markers, and incomplete words. Each segment was adjusted to roughly correspond to a sentence. The edited texts are being published online alongside the corresponding audio recordings.

Additionally, we have been converting the ELAN (.eaf) files containing plain transcriptions and Spanish translations into .flex format and uploading them into FieldWorks, where morphological glossing and English translations are added. We chose FieldWorks (FLEx) over ELAN for morphological annotation because it allows for automatic parsing and helps ensure better consistency in glosses. At present, a fully annotated subcollection is available upon request.

6. Results.

6.1. PROJECT'S OUTCOMES. The project is currently reaching completion, with significant progress already achieved. The corpus contains over eight hours of lightly edited (see above) audio recordings in Patzún Kaqchikel, listed in Table 1. Each recording is accompanied by an ELAN (.eaf) annotation file that includes a transcription, a translation in Spanish, and additional tiers for cultural commentary and standardized orthography. These files have been reviewed by the PIs and are ready for glossing.

Recording title/topic	Length
Recipes	2:52:41
Traditional embroidery	0:17:29
Marimba	0:16:18
Traditional medicine and herbs	0:17:29
Maternity	0:30:45
Kaqchikel children, education	0:27:01
Kaqchikel evening school	0:12:17
Mayan center (Centro de Estudio Maya)	0:29:31
Women's Cooperative Aj Su'm	0:38:52
Local business	0:20:37

Mayan religion	0:12:41
Catholicism in Patzún	0:53:40
the Guatemalan Civil War	0:19:47
Iximche'	0:20:04

Table 1. Recordings in the Patzún Kaqchikel corpus

The collection is currently available on demand as well as through a separate community-oriented web page (Google Sites). It is also uploaded online as part of the Kaqchikel corpus developed by Ryan Bennett and Robert Henderson (2022) using LingView. The original .wav audio files were converted to .mp3 to save space; the online corpus contains the .mp3 files together with the corresponding .eaf files, and on the web page the texts are embedded. The community web page will be published within the course of 2026; the link will be shared via personal communication with members of the community, in particular, the members of the Co-operative that we have been in touch with.

A sub-collection of 3,000 tokens is also available as a FieldWorks project with morphological glossing and English translation. We will shortly expand it to 5,000 tokens, which will already make it suitable for training machine-learning NLP models; see Moeller & Liu (2024) on 1,500-6,000 (2,000-4,000) tokens being the optimal minimum amount of annotated data to train a model. When fully glossed, we expect the collection to reach about 50,000 tokens.

6.2. PROJECT'S IMPACT. Although still ongoing, the project will benfit both the Kaqchikel speaker community and the broader linguistic community.

For the Kaqchikel community, an open-access collection of narratives can help preserve linguistic and cultural practices. Making the corpus available online will also enhance the prestige of the language, because the corpus will validate its importance in a novel manner. The culture and history of Patzún have already attracted some attention; consider, for example, the online resource Qanatab'äl – Escribamos la historia de Patzún. Adding Kaqchikel narratives to the existing stories in Spanish is a step forward in community-driven cultural preservation. The project also carries personal significance. Several interviewees passed away during the pandemic, and the recordings of their voices have become irreplaceable. We hope that they will provide comfort to their families and be cherished as a living legacy by the community, including its younger members.

From a research perspective, the corpus is a valuable resource for linguistic, anthropological, and cultural studies, enriching Kaqchikel representation in theoretical linguistics, language typology, and sociolinguistics. Its breadth of topics and diverse speaker profiles make it useful to researchers working on micro-variation and dialectal comparison, code-switching, discourse particles, and Kaqchikel culture and history, along with any other research questions that may emerge in the future.

7. Concluding remarks.

7.1. SUMMARY. In this paper, we have presented a language documentation project on Patzún Kaqchikel that was substantially reshaped by the constraints of the COVID-19 pandemic. Originally conceived as a community-oriented initiative to produce a trilingual book of recipes and oral histories, the project shifted toward developing an open-access corpus of spoken Kaqchikel, due to the team's inability to conduct in-person fieldwork. The paper has discussed the major challenges that the project faced and the adjustments that were made to overcome them.

The resulting collection consists of over eight hours of narratives recorded from native speakers of Kaqchikel from diverse social and occupational backgrounds. These materials are being made available in two complementary formats: (i) a community-oriented web page with recordings and accessible transcripts in Kaqchikel and in Spanish, and (ii) an online open-access annotated corpus with time-aligned transcriptions and translations in Spanish; a sub-corpus with morphological glossing and translations in English is available on demand.

7.2. COMPARISON TO OTHER PROJECTS. Our experience of having to adjust to the conditions of the COVID-19 pandemic is not unique; many other language-documentation projects were likewise significantly affected. Two adjustment strategies frequently reported in the literature are (i) relying on previously established community relationships (Griscom 2020), and (ii) focusing on already collected materials (Bowern 2020). We adopted both strategies, albeit with certain limitations. In 2020, when fieldwork became impossible, we were still relatively new to the community: Burukina had been conducting fieldwork in Patzún every summer since 2017, and Pleshak, since 2018. While we stayed in contact with members of our host families through social media, we did not have sufficient time to forge closer friendships and mostly relied on our connections to the Cooperative and the Guatemala Field Station. We were fortunate to record several interviews in 2019 that could be used for ongoing work, but, as acknowledged above, the material was not enough to prepare a full book manuscript.

Our experience differs from many other pandemic-affected projects in the direction of methodological shift. Projects that originally focused on academic research often became community-centered when remote collaboration demanded greater reliance on local researchers (e.g., Singer 2020). Our project was originally designed as community-oriented; however, the inability to return to Patzún reduced collaborative engagement and pushed the project toward a more traditional corpus-development workflow. We hope that, despite these differences between other documentation projects affected by the pandemic and ours, we can meet in the "golden middle", ultimately addressing both community and academic needs.

Some researchers forced to shift to remote asynchronous work report advantages of a flexible schedule (Morey 2020; Williams et al. 2021), with consultants and research assistants being able to incorporate linguistic work more easily into their lives (Williams et al. 2021). In contrast, we experienced the opposite: community members prioritized their families and work, especially given that Guatemala was severely affected by the pandemic (Martinez-Folgar et al. 2021, Díaz-Bonilla et al. 2022, Mercadal 2022, i.a.). We do not claim that efficient remote work is impossible, but we point out that switching to online does not always simplify the workflow. For our project, while the community appreciated our efforts and interest in their language and culture, they did not seek active participation in or control over the project, and the pandemic and subsequent lockdowns increased the distance between the community and the PIs.

In conclusion, this project makes a case that extends beyond its immediate findings: that community-centered research and academic rigor are not in tension but complementary. Crisis conditions do not suspend scholarly standards—they test and ultimately strengthen them, provided researchers remain responsive to the communities whose knowledge they document.

Corpora

Bennett, Ryan & Robert Henderson. 2022. Online corpus of Kaqchikel texts (Mayan, Guatemala). https://bkeej.github.io/TextosKaqchikeles/#/ (accessed 27 March 2026)

References

Bowern, Claire. 2020. Yale grammar boot camps. A talk presented online at Linguistic fieldwork: Working with communities at a distance, organized by Abralin ao vivo. 29 July 2020. https://youtu.be/egnbg4KuX5E (accessed 27 March 2026)

Díaz-Bonilla, Eugenio, Luis Flores, Valeria Piñeiro & Miriam Centurión. 2022. Guatemala: The impact of COVID-19 and other shocks, and policy implications: Final report. LAC Working Paper 28. Washington, DC: International Food Policy Research Institute. https://doi.org/10.2499/p15738coll2.136358

Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2022. *Ethnologue: Languages of the world*. 25th edition. Dallas, TX: SIL International. http://www.ethnologue.com.

Griscom, Richard E. 2020. Remote linguistic elicitation methods. The Endangered Languages Archive, 25 June 2020. https://blogs.soas.ac.uk/elar/2020/06/25/remote-linguistic-elicitation-methods/ (accessed 27 March 2026)

Heinze, Ivonne. L. 2004. *Kaqchikel and Spanish language contact: The case of bilingual Mayan children*. Lawrence: University of Kansas dissertation.

Kashkin, Egor V. 2020. Polevye issledovanija vostočnyx gornomarijskix govorov: korpus tekstov [Field studies of the Eastern Hill Mari idioms: A text corpus]. In *XIX Ignatjevskie čtenija. Materialy dokladov i vystuplenij na ežegodnoj regional'noj naučno-praktičeskoj konferencii "Gornye mari v kontekste istorii i kul'tury", posv'aščennoj Godu teatra v Rossii [XIX Ignatjev Readings. Proceedings of the yearly regional scientific conference "Hill Mari in the context of history and culture," dedicated to the Year of Theater in Russia]*. 3–15. Yoshkar-Ola: MarNIIJLI.

Kaufman, Terrence. 1974. *Idiomas de Mesoamérica*. Guatemala: Seminario de Integración Social.

Martinez-Folgar, Kevin, Diego Alburez-Gutierrez, Alejandra Paniagua-Avila, Manuel Ramirez-Zea & Usama Bilal. 2021. Excess mortality during the COVID-19 pandemic in Guatemala. *American Journal of Public Health* 111(10). 1839–1846. https://doi.org/10.2105/AJPH.2021.306452

Mercadal, Trudy. 2022. Three challenges facing Guatemala's COVID-19 crisis: Mobility, violence and governance. In Stanley D. Brunn & Donna Gilbreath (eds.), *COVID-19 and a world of ad hoc geographies*. 249–267. Cham: Springer. https://doi.org/10.1007/978-3-030-94350-9_16

Moeller, Sarah & Zoey Liu. 2024. Leveraging AI for language documentation. A talk presented at the 9th International Conference on Language Documentation & Conservation. 4–9 March 2025. Honolulu, HI.

Morey, Stephen. 2020. Linguistic fieldwork: Working with communities at a distance. A talk presented online at Linguistic fieldwork: Working with communities at a distance, organized by Abralin ao vivo. 29 July 2020. https://youtu.be/egnbg4KuX5E (accessed 27 March 2026)

Moseley, Christopher (ed.). 2010. *Atlas of the world's languages in danger*. 3rd edition. Paris: UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000187026.

Patal Majzul, Filiberto. 2007. *Rusoltzil ri Kaqchikel: Diccionario bilingüe estándar Kaqchikel illustrado*. Guatemala: OKMA.

Patal Majzul, Filiberto, Lolmay Pedro García Matzar & Carmelina Ixchel Espantzay Serech. 2000. *Rujunamaxik ri Kaqchikel chi': Variación dialectal en Kaqchikel*. Guatemala: Editorial Cholsamaj.

Polinsky, Maria. 2019. Field stations for linguistic research: A blueprint of a sustainable model. *Language* 95(2). e327–e338. https://doi.org/10.1353/lan.2019.0045.

Richards, Michael. 2003. *Atlas Lingüístico de Guatemala*. Instituto de Lingüístico y Educación de la Universidad Rafael Landívar.
Romero, Sergio. 2015. *Language and ethnicity among the K'ichee' Maya*. Salt Lake City: University of Utah Press.
Romero, Sergio. 2017. The labyrinth of diversity: The sociolinguistics of Mayan languages. In Judith Aissen, Nora C. England & Roberto Zavala Maldonado (eds.), *The Mayan languages*. 379–400. London: Routledge.
Singer, Ruth. 2020. Collaborative ways of researching multilingualism: Coordinating activities at a distance. A talk presented online at Linguistic fieldwork: Working with communities at a distance, organized by Abralin ao vivo. 29 July 2020. https://youtu.be/egnbg4KuX5E (accessed 27 March 2026)
Williams, Nicholas., Wilson D.L. Silva, Laura McPherson & Jeff Good. 2021. COVID-19 and documentary linguistics: Some ways forward. *Language Documentation and Description* 20. 359–377. https://doi.org/10.25894/ldd57