FLORES+ Mayas: Generating Textual Resources to Foster the Development of Language Technologies for Mayan Languages

Andrés Lou, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena

Transducens Research Group, Universitat d'Alacant, Spain
{and_lou,japerez,fsanchez,miquel.espla,vm.sanchez}@ua.es

Abstract

A significant percentage of the population of Guatemala and Mexico belongs to various Mayan indigenous communities, for whom language barriers lead to social, economic, and digital exclusion. The Mayan languages spoken by these communities remain severely underrepresented in terms of digital resources, which prevents them from leveraging the latest advances in artificial intelligence. This project addresses that problem by means of: 1) the digitisation and release of multiple printed linguistic resources; 2) the development of a high-quality parallel machine translation (MT) evaluation corpus for six Mayan languages. In doing so, we are paving the way for the development of MT systems that will facilitate the access for Mayan speakers to essential services such as healthcare or legal aid. The resources are produced with the participation of indigenous communities: native speakers provide the necessary translation services, QA, and linguistic expertise. The project is funded by the Google Academic Research Awards and carried out in collaboration with the Proyecto Lingüístico Francisco Marroquín Foundation in Guatemala.

1 Introduction

Recent advances in natural language processing (NLP) and artificial intelligence (AI) come with the caveat of needing a sufficiently large amount of data. These data are far from being available for the indigenous Mayan languages, which cover a historical region comprising Guatemala, Belize, and southern Mexico. Our project "Generating Textual Resources to Foster the Development of Language Technologies for Mayan Languages" aims at creating textual resources for Mayan languages as a first step for their language communities to take advantage of recent AI advances. Our two concrete goals are: digitising multiple printed linguistic resources and releasing them as data artefacts; and creating a parallel corpus, FLORES+ Mayas, to establish a high-quality benchmark for the development and evaluation of machine translation (MT) systems for some Mayan languages. This corpus will be obtained by translating the Spanish side of the FLORES+ corpus[1] (Goyal et al., 2022; NLLB Team et al., 2022), a de-facto standard for low-resource MT. Recognising the importance of indigenous participation, summarised in the epigram "Nothing about us without us", the Universitat d'Alacant has signed a formal collaboration agreement with the Proyecto Lingüístico Francisco Marroquín Foundation[2] (FPLFM), an indigenous, non-government organization from Guatemala with a long history in terms of language documentation and preservation. They provide us with the physical media to be digitised and are also our main liaison with the Mayan translation community in Guatemala.[3]

The project is funded by the Google Academic Research Award (GARA), a program open to professors at degree-granting institutions who are conducting research in the field of technology and computing.[4] The project started at the beginning of 2025 and plans to be completed by year-end. All developments will be hosted on https://github.com/transducens/floresmayas under open licenses: CC BY-SA 4.0 for FLORES+ Mayas and a yet-to-be-decided license for digitised resources.

2 Description of the proposed work

2.1 Digitisation of Mayan linguistic resources

FPLFM and INALI provided our research group with a number of physical copies of several text

[1] https://oldi.org
[2] https://plfm.org
[3] There is also an ongoing collaboration with the National Institute of Indigenous Languages (INALI) in Mexico.
[4] https://research.google/programs-and-events/google-academic-research-awards

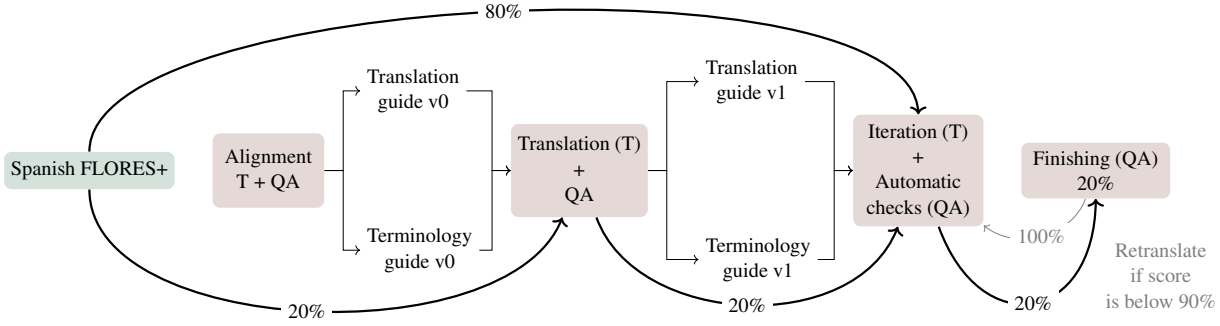

Figure 1: Workflow for developing FLORES+ Mayas (see main text for details).

sources, namely Mayan-Spanish bilingual dictionaries, grammar books, and bilingual narratives, in the following languages: Awakatek, Chuj, Jakaltek, Kaqchikel, Mam, Q'eqchi', Tz'utujil, K'iche', and Yucatec Mayan.[5] We digitised approximately 30 000 bilingual entries, 135 200 words of monolingual grammar descriptions, and 188 500 words from narratives. The digitisation was supported by the Miguel de Cervantes Virtual Library,[6] one of the largest open repositories of digitised Spanish-language historical texts. We are currently exploring the use of pure OCR engines, such as Tesseract (Smith, 2007), the use of multimodal LLMs, such as Google Gemini, and the combination thereof. The resulting documents will be further curated in order to be released as textual resources that may be used to train MT systems or language models (Tanzer et al., 2024). Given the historical lack of standardization of Mayan orthography (López Raquec, 1989), we will transcribe each resource into the corresponding modern standards.

2.2 FLORES+ Mayas

The development of FLORES+ Mayas involves the manual translation of the Spanish dev and devtest fractions (around 2 000 sentences and 50 000 words) of the the FLORES+ corpus into K'iche', Kaqchikel, Ixil, Mam, Q'anjob'al, and Q'eqchi'; these languages were selected on the criteria of availability of resources, number of speakers, and language taxonomy. We plan to organize an on-site translation task in Guatemala in the context of our agreement with FPLFM, where teams of indigenous native speakers will translate the corpus sentences. We will follow the methodology described by NLLB Team et al. (2022) (see Figure 1) consisting of the following stages: 1) **Alignment**: teams get acquainted with the nature of the data and the task, e.g. preparing lists of neologisms, discussing spelling and orthography, etc. 2) **Translation**: 150 sentences are translated for each language and sent to the QA analyst for review and feedback. 3) **Iteration**: Adjustments based on feedback are made and then the full translation of all the sentences is performed. 4) **Completion**: A sample (20%) of the translated sentences is assessed by the QA analyst, which will determine whether the quality is acceptable or retranslation is needed.

[5]The parallel data currently available ranges from a few thousand words (Awakatek) to a few million (Yucatec Mayan).
[6]https://www.cervantesvirtual.com

Acknowledgments

Project funded by Google through the 2024 Google Academic Research Awards program (Society-Centered AI RFP).

References

Naman Goyal et al. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Margarita López Raquec. 1989. *Acerca de los alfabetos para escribir los idiomas mayas de Guatemala: proyecto lingüístico Francisco Marroquín, Antigua Guatemala, Julio 1988*. Ministerio de Cultura y Deportes.

NLLB Team et al. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Garrett Tanzer et al. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.