Curated Datasets and Neural Models for Machine Translation of Informal Registers between Mayan and Spanish Vernaculars

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

Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
E-03690 Sant Vicent del Raspeig, Spain

and_lou@ua.es, {japerez,fsanchez,vmsanchez}@dlsi.ua.es

Abstract

The Mayan languages comprise a language family with an ancient history, millions of speakers, and immense cultural value, that, nevertheless, remains severely underrepresented in terms of resources and global exposure. In this paper we develop, curate, and publicly release a set of corpora in several Mayan languages spoken in Guatemala and Southern Mexico, which we call MayanV. The datasets are parallel with Spanish, the dominant language of the region, and are taken from official native sources focused on representing informal, day-to-day, and non-domain-specific language. As such, and according to our dialectometric analysis, they differ in register from most other available resources. Additionally, we present neural machine translation models, trained on as many resources and Mayan languages as possible, and evaluated exclusively on our datasets. We observe lexical divergences between the dialects of Spanish in our resources and the more widespread written standard of Spanish, and that resources other than the ones we present do not seem to improve translation performance, indicating that many such resources may not accurately capture common, real-life language usage. The MayanV dataset is available at https://github.com/transducens/mayanv.

1 Introduction

The Mayan language family is spoken in an area covering the modern states of Guatemala, Belize, and Southern Mexico (Law, 2014, p. 81). It consists of around 30 languages grouped into five or six major sub-groups, depending on the source (Campbell and Kaufman, 1985; Law, 2014). Most subgroups and most speakers are found today in Guatemala, where between 40-60% of the population are native speakers (Instituto Nacional de Estadística, 2018; England, 2003). Mayan languages are relatively healthy, but their presence online and on the global scene in general is almost non-existent. In effect, Mayan languages, despite the total number of speakers, are considered to be somewhat in decline: according to Richards and Macario (2003), only around half the population of ethnic Mayas are Mayan speakers, and the languages are associated in many social contexts to backwardness, ignorance and poverty (England, 2003).

Figure 1: Sample of the ancient Mayan script, reading *b'alam*, "jaguar", using a combination of the logogram and the syllabogram. Attribution: Goran tek-en under license CC BY-SA 4.0.

To begin addressing the problem of lack of representation in the digital realm, and increase access to modern technology and information sources for indigenous communities, we seek to develop neural machine translation (NMT) (Koehn, 2020) systems. To train and evaluate them, it is first necessary to produce and curate corpora of all the languages involved. In this paper, we present MayanV, a series of curated parallel corpora between various Mayan languages and Spanish. The language register in these datasets is informal, familial, and non-domain-specific, which best reflects the most common use of the languages involved. In general, online resources for building working NMT models are very scarce: the two greatest sources for parallel texts are the Bible, whose overly formal and potentially archaic language does not reflect most modern use cases, and the parallel texts of the Jehova's Witnesses website (jw.org), whose language, though much closer to modern usage, is still somewhat divorced from the common, day-to-day activities carried out by most Mayan speakers,

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2838–2850

as our dialectometric analysis suggests. Outside these two sources, bilingual and monolingual texts longer than a few thousand sentences are rare and not suited for processing, existing mostly as human-readable PDF files. Because of such scarcity of parallel resources for any Mayan language, especially those with just a few thousand, or even a few hundred, speakers, we use the parallel corpora we have built to train and evaluate a number of bilingual and multilingual NMT systems; in particular, multilingual systems have proven effective when dealing with low-resource and underrepresented languages (Lakew et al., 2018).

While other notable efforts in NMT of low-resource and endangered languages have been carried out recently, such as the No Language Left Behind (NLLB) (Team et al., 2022) and MADLAD projects (Kudugunta et al., 2023), these include little to no focus on Mayan languages. Our paper, in contrast, focuses on a more formal introduction of the Mayan languages to the larger natural language processing (NLP) community and on the presentation, curating, and release of parallel datasets that may be used for benchmarking future translation endeavours. The corpora are available at https://github.com/transducens/mayanv/.

The rest of the paper is organised as follows. The next section offers a historical and linguistic overview of the Mayan languages motivated by their relative obscurity amongst the NLP community. Section 3 then presents the related work in the field. Section 4 presents in detail the extraction and curating of resources for dataset creation and model training, including a dialectometric analysis by which we characterise the dialectal and register divergence between the Spanish found in MayanV and the more standard variety of jw.org. Afterwards, Section 5 describes and evaluates the NMT systems we have built. The paper ends with final remarks and a description of potential future research directions.

2 Overview of the Mayan Languages

The oldest attested Mayan language, referred to as Classic Maya, dates back to ca 300 BC. Written in the the Mayan script, it belongs to a tradition corresponding to the few instances in human history in which writing was independently invented, along with the systems developed in Ancient Egypt, Sumer, and Ancient China (Fagan, 1996, p 762). Figure 1 shows an example of the script. The

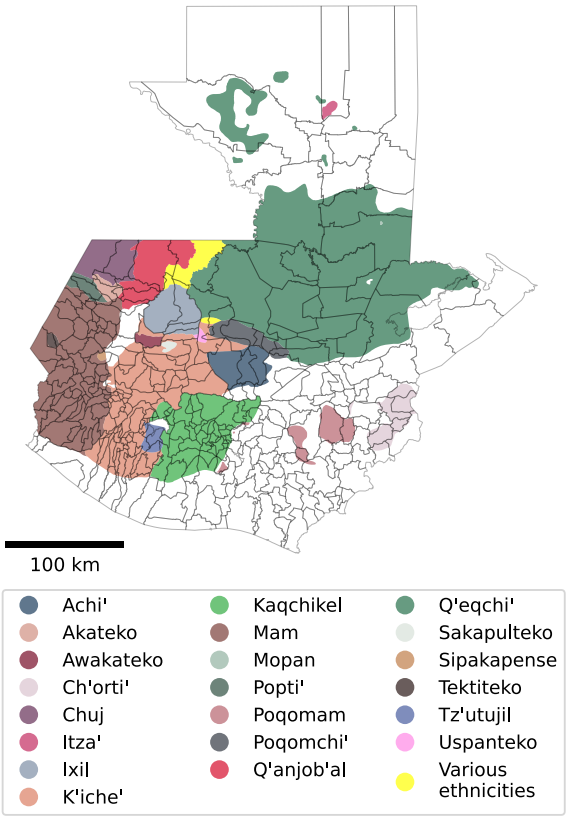

Figure 2: The Mayan linguistic communities of Guatemala

Mayan script is logosyllabic: glyphs may act either as logograms, representing a complete semantic unit, or syllabograms, representing a syllable (Coe and Van Stone, 2016). Modern Mayan languages are written using the Latin script, with additional diacritics used to denote features such as vowel length and glottalisation, amongst others.

Much variation exists in terms of breadth and number of speakers. The three most spoken languages —K'iche', Yucatec Mayan, and Q'eqchi— all have between a million and half a million native speakers, and are spoken across different geographical areas in Guatemala and Mexico (Richards and Macario, 2003; Law, 2014). In contrast, there are some languages with very few speakers and in imminent danger of language death, as is the case of Itza', which is only spoken by elderly adults and has fewer than 1 000 native speakers (Instituto Nacional de Estadística, 2018; Eberhard et al., 2021). However, despite their size in terms of speakers and geographical extent, as shown in Figure 2, Mayan languages are usually not given an official status in their respective countries; their use is widespread in daily activities and familial environments, but they

are nearly non-existent in matters of governance, education, mass media, and healthcare (Romero, 2017). For example, in Guatemala, during Covid-19 pandemic, the official online portal to register for the government-sponsored vaccination program was only accessible in Spanish (España, 2021); to this day it continues to lack any Mayan language version. Even when official support is said to exist, the actual implementation of educational initiatives remain deficient, as is the case with incipient government-sponsored programs to teach Yucatec Mayan in schools (Bote Tun, 2023).

Bilingualism with Spanish is very common, though not uniform across geographical, population and gender lines: isolated communities often present a high number of monolingual speakers, and men usually exhibit higher proficiency in bilingualism (Bennett et al., 2016; Romero, 2017; Richards and Macario, 2003). Language shift, whereby a Mayan language is displaced by either Spanish or, less commonly, another Mayan language (Bennett et al., 2016; Romero, 2012), occurs more quickly in urban environments, where a lingua franca is expected to be used.

Literacy is overall low as a result of decades of policy where using native languages in schools was discouraged or even punished (French, 2010). As a result, many Mayan speakers regard the orthography of their own languages as more difficult and inaccessible than that of Spanish. Change was slowly brought about with the introduction of bilingual educational programs in the early 1990s, which finally led to a continued effort to revitalise the role of written Mayan languages. The Guatemalan Academy of Mayan Languages (ALMG), established in 1990, plays an important role in the standardisation of both the orthography of the different *linguistic communities* and their corresponding spoken languages, while also engaging in literacy and publication efforts.[1] A similar role is played by the Proyecto Lingüístico Francisco Marroquín Foundation (PLFM),[2] also in Guatemala, and National Institute of Indigenous Languages (INALI), established in 2003, in Mexico,[3] both of which have published several grammars.

Mayan languages exhibit a high degree of dialectal variation (Romero, 2017). While the most important reason for this is natural language change and historical innovation, the sprachbund of Mayan languages in Guatemala and Southern Mexico has resulted in much linguistic exchange in the forms of loanwords and calques, usually manifesting as the influence of a language with more speakers and political leverage over another with fewer speakers and less influence. Dialectal divergence within the same language is also attested, as is the case with the dichotomies of the Western and Eastern, and Standard and Lowland dialects of Q'eqchi' (DeChicchis II, 1989; Romero, 2012). Additionally, the politics of identity play a key role in delimiting the difference between a language and a dialect in the mind of their speakers, as seen in the cases of Achi, Akatek, and Chalchitek, which are sometimes considered dialects of K'iche', Q'anjob'al, and Awakatek, respectively (Bennett et al., 2016). In general, Mayan languages exhibit limited mutual intelligibility, and code-switching with Spanish and other Mayan languages in areas of high contact is common (Little, 2009).

Despite their relative obscurity in the field of NLP, Mayan languages are well studied and documented in matters of historical linguistics, morphosyntax, phonology, and semantics, as seen in extensive works such as those by Bennett et al. (2016), Bennett (2016), Coon (2016), Henderson (2016), Polian (2017) and many others.

3 Related Work

African languages exist in a similar, albeit superlative, situation to that of Indigenous languages in the Americas. Thousands of African languages exist across the continent, boasting millions of speakers and serving as an important symbol of culture and cultural exchange; nevertheless, they are poorly represented in NLP applications. The Masakhane project (Orife et al., 2020) seeks to address this situation by fostering a community of researchers and non-researchers alike dedicated to advancing the development of NLP for African languages. Martinus and Abbott (2019) describes a number of challenges the African NLP community faces: Little official support for African indigenous languages; lack of resources for any kind of NLP task, and when those resources exist, they are hard to find; lack of benchmarks; and low reproducibility.

Mayan languages, and indigenous languages in general, face the same issues. Nevertheless, the work on NLP of Indigenous languages of the Americas continues (Mager et al., 2023). Of note is the

[1] https://www.almg.org.gt/nosotros/historia

[2] https://plfm.org/quienes-somos/historia

[3] https://site.inali.gob.mx/Micrositios/normas/index.html

work by Tyers and Henderson (2021) and Tyers and Howell (2021), who focus specifically on K'iche' and develop an annotated corpus for morphosyntactic structure and perform a survey of part-of-speech tagging methods respectively, and also Pugh et al. (2023) who work on a finite-state transducer for performing morphological analysis on Yucatec Maya. Similarly to our work, Oncevay (2021) presents a multilingual NMT system, including both Spanish-to-many and many-to-Spanish models for Aymara, Ashaninka, Quechua, and Shipibo-Konibo, all indigenous languages spoken in Peru; in general, efforts to bring modern MT into the realm of endangered indigenous languages are gaining traction, with work focused on Nahuatl, Otomi, Guarani, Quechua, and many other prominent indigenous languages from countries where there exist a sizeable population of indigenous peoples (Mager et al., 2021; Parida et al., 2021; Knowles et al., 2021; Zheng et al., 2021; Vázquez et al., 2021). Crucially, however, Mayan languages are nearly non-existent in these endeavours.

The introduction of multilingual and cross-lingual NMT (Johnson et al., 2017; Conneau and Lample, 2019), along with their application in low-resource scenarios (Lakew et al., 2018; Karakanta et al., 2017; Madaan and Sadat, 2020) was of vital importance in the effort to improve NMT in languages with otherwise small or almost non-existent written bodies of work. Recent work on low-resource languages has been put forward by Meta's NLLB project (Team et al., 2022), a contribution that includes several new benchmarks, including the FLORES+ dataset[4] (based on FLORES-200, an update of FLORES-101 (Goyal et al., 2021)), and a state-of-the-art NMT translation model, called NLLB-200, focusing on underrepresented and low-resource languages, though, unfortunately, not a single Mayan language was included in their efforts. In contrast, Google's MADLAD-400 dataset, along with its accompanying translation model (Kudugunta et al., 2023), does include a number of Mayan languages in the form of monolingual corpora originating from jw.org and Bible sources.

4 Development and Curation of MayanV

To develop the MayanV corpora, we manually crawled, extracted, and cleaned a number of online resources, mostly published by the ALMG, except for the Tzeltal dictionary (Polian, 2018); these extracted resources, which we collectively call the Mayan Vocabularies following the naming convention laid out by the ALMG, include the following languages: Achi (Teletor Velásquez et al., 2016), Awakatek (Academia de Lenguas Mayas de Guatemala, 2006), Chuj (García Mendoza et al., 2003), Itza' (Academia de Lenguas Mayas de Guatemala, 2020), Ixil (Laynez Ayay et al., 2018), Q'eqchi' (Caal Ixim et al., 2004), Q'anjob'al (Pablo Escobar et al., 2003), Mam (López Mejía et al., 2004), Poqomam (Conguache Coj et al., 2001), Poqomchi' (Morán Mus et al., 2001), K'iche' (Pérez Medrano and Delgado, 2010), Sipakapense (Tema Bautista et al., 2017), Tektitek (Méndez Pérez et al., 2018), and Tz'utujil (Ixcaya Ratzam et al., 2019).

4.1 Extracting the Mayan Vocabularies

The corpora in the Mayan Vocabularies can all be described as lists of entries, where each entry contains a Mayan word, its corresponding translation into Spanish, at least one example of the usage of the word in the Mayan language, and the corresponding translations into Spanish of such usages. Fourteen out of the fifteen corpora were only accessible as densely-formatted PDFs, the exception being the Tzeltal dictionary. See Figure 3 for an example of the format and Table 1 for a breakdown of the languages involved and the sizes of their respective corpus. Using `pdfplumber`,[5] the process by which we extracted the corpora can be summarised as follows: Using a heuristical approach involving the relative position of non-textual elements, such as margin delimiters or decorations, and that of typographically significant elements, e.g. the left-most bold-font character, each entry in the document was added to a list of bounding boxes and extracted. Words were then counted using white space, and sentences were split using punctuation and typography. Typically, each entry uses a particular font style for Mayan text and another different font style for Spanish text, as shown in Figure 3a. We were only interested in extracting full sentences, which were typically delimited by style and punctuation, and were often found in the latter parts of an entry. However, some entries included more than one example of usage, meaning than additional, oftentimes manual, inspection was required to extract the parallel text and avoid mixing up the languages.

Using the Mayan Vocabularies, we developed a

[4] https://github.com/openlanguagedata/flores

[5] https://github.com/jsvine/pdfplumber

Ab'lil aatin. *Extranjerismo.* Sa' li qaatinob'aal yo xch'ikb'al rib' naab'al ab'lil aatin. *En nuestro idioma se están introduciendo muchas palabras extranjeras.*

(a) Original text, formatting, and layout.

```
Sa' li qaatinob'aal yo xch'ikb'al
rib'naab'al ab'lil aatin.
En nuestro idioma se están introduciendo
muchas palabras extranjeras
```

(b) Resulting parallel sentences.

Figure 3: (3a) Entry in the Q'eqchi' corpus of MayanV. The Q'eqchi' term is in bold font; the first set of italics is the Spanish translation, "loanword"; the regular text is a usage example of the term; the second set of italics is the Spanish translation of the example. (3b) Extracted Q'eqchi' sentence and its Spanish translation: "There are many loanwords being introduced into our language."

number of benchmarks intended to encourage other researchers to join the effort of developing NMT models for Mayan languages. We have named the resulting dataset MayanV. The corpora that comprise MayanV are part of the language and spelling standardisation efforts carried out by the ALMG. As such, in spite of the documented dialectal variation of some of the listed languages, we consider them good representations of modern and widespread language use.

All mined PDFs are freely available for download at the ALMG's website, though there is considerable variation amongst them in terms of length, layout, typography, and content. This not only makes the extraction task laborious but also means that copy-editing and encoding errors are common in the original documents. Indeed, post-extraction editing was necessary in several instances. For example, in the Mam (mam) corpus, the phonemic consonant /ʃ/ is represented with the wrong character ("õ" instead of the digraph "xh") throughout the whole text, a fact that only came to light after thorough cross-examination with other corpora.

4.2 Spanish Dialectometry of MayanV

The most salient trait of MayanV, standing in contrast to many of the available resources for MT, is the dialect and register in which both source and target languages are written. Much of the language in these texts is informal, day-to-day, and non-specific to any domain. Additionally, we note that much of the Spanish in the parallel texts is vernacular to Guatemala and Southern Mexico. As already mentioned, this is one of the most important aspects of MayanV as training and testing resources, as they reflect the most common use case amongst the marginalised indigenous minorities of Guatemala.

ISO	Language	Words (Mayan)	Words (es)	Sentences
acr	Achi	6 994	7 657	1 343
agu	Awakatec	7 325	9 700	1 930
cac	Chuj	9 398	10 916	2 299
itz	Itza'	6 069	7 512	1 539
ixl	Ixil	10 888	12 137	2 325
kek	Q'eqchi'	18 529	21 835	4 133
kjb	Q'anjob'al	18 035	18 238	3 014
mam	Mam	15 453	19 117	3 093
poc	Poqomam	18 039	21 744	3 583
poh	Poqomchi'	6 479	7 149	1 787
quc	K'iche'	14 468	15 474	2 632
qum	Sipakapense	9 780	9 328	1 356
ttc	Tektitek	23 571	24 896	4 022
tzh	Tzeltal	103 309	128 659	19 846
tzj	Tz'utujil	12 283	11 404	2 519

Table 1: The 15 corpora of MayanV curated for our work. Size in terms of parallel sentences and words of the parallel corpora extracted from them. Sources for each corpus are discussed in the main text.

Following the example of Donoso and Sánchez (2017), we use the relative frequencies of the synonyms of a set of curated concepts taken from the Varilex project (Ueda and Ruiz Tinoco, 2003) to compute metrics that yield a sense of the lexical variation between contrasting dialects. As an example, the concept "earthquake" is materialised in the Spanish words *movimiento*, *movimiento sísmico*, *remezón*, *sacudida*, *seísmo*, *temblor*, *temblor de tierra*, and *terremoto*. We seek to compare the dialects of Spanish from MayanV and the Spanish used in jw.org.

We represent each concept as a vector of the frequencies of each synonym. We then compute the average cosine distances amongst the Spanish texts of MayanV, and, when possible, the distance between the Spanish in MayanV and the jw.org corpora. For any two synonym vectors $\mathbf{s}_i, \mathbf{s}_j$ we compute the cosine distance as

$$1 - \frac{\mathbf{s}_i \cdot \mathbf{s}_j}{\|\mathbf{s}_i\| \|\mathbf{s}_j\|}.$$

We only include vectors with at least one non-zero component, and only compare concepts appearing in both corpora. Results are shown in Table 2. Lower values indicate dialectal proximity, though some figures might be less indicative than

	n_{jw}	jw	acr	agu	cac	itz	ixl	kek	kjb	mam	poc	poh	quc	qum	ttc	tzh	tzj
acr	-	-	-	17	17	15	16	18	15	18	17	14	17	16	16	21	18
agu	-	-	0.175	-	21	16	16	21	18	19	19	15	20	20	18	21	19
cac	-	-	0.320	0.284	-	21	20	30	20	22	24	14	23	22	25	27	23
itz	-	-	0.241	0.252	0.295	-	15	23	14	17	18	14	16	16	17	19	20
ixl	-	-	0.227	0.202	0.293	0.238	-	19	16	18	21	13	18	17	21	22	19
kek	28	0.428	0.281	0.248	0.185	0.215	0.258	-	21	24	22	15	23	23	24	27	25
kjb	-	-	0.193	0.150	0.259	0.251	0.172	0.221	-	20	20	14	18	17	16	21	17
mam	23	0.402	0.188	0.164	0.269	0.224	0.181	0.212	0.115	-	20	15	21	19	18	24	20
poc	-	-	0.272	0.219	0.301	0.287	0.203	0.262	0.205	0.210	-	15	22	18	23	28	21
poh	12	0.313	0.168	0.142	0.274	0.194	0.180	0.232	0.158	0.149	0.184	-	14	15	14	17	14
quc	24	0.325	0.265	0.245	0.267	0.256	0.227	0.205	0.235	0.188	0.241	0.196	-	20	23	24	21
qum	-	-	0.293	0.200	0.303	0.266	0.264	0.251	0.254	0.204	0.316	0.201	0.184	-	21	24	21
ttc	-	-	0.354	0.271	0.316	0.299	0.219	0.233	0.291	0.277	0.256	0.245	0.216	0.294	-	26	19
tzh	38	0.387	0.359	0.264	0.338	0.364	0.283	0.304	0.265	0.242	0.252	0.246	0.281	0.319	0.292	-	24
tzj	-	-	0.231	0.227	0.285	0.223	0.190	0.211	0.214	0.186	0.240	0.191	0.191	0.228	0.291	0.310	-

Table 2: Average cosine distance and number of overlapping concepts between each of the Spanish texts in MayanV, including the dialect from jw.org when available. Columns n_{jw} and jw denote the number of overlapping concepts and the average lexical distance between the Spanish in the corresponding MayanV corpus and the Spanish from jw.org. The upper diagonal of the table indicates the number of overlapping concepts between the Spanish of the respective MayanV corpora, while the lower diagonal indicates their lexical distance (average cosine distance). Lower values indicate dialectal proximity.

others given the low number of overlapping concepts for any given pair of languages in general. Nevertheless, we observe evidence of considerable variation in the distances between these, possibly reflecting the lack of cross-linguistic regulation during the production of the documents from which MayanV was developed. We also observe a noticeable divergence with the dialect of jw.org, which is empirically closer to a more widespread written standard.

The characterisation of the Spanish dialects involved in our task is of particular importance, given that the language acts as the most widespread common tongue in the region. The divergence of Spanish into several regional and mutually intelligible dialects is well-attested, even though the Spanish dialect of Guatemala remains somewhat understudied. Nonetheless, there have been important efforts towards its documentation (Pato, 2023; Kotenyatkina, 2019). We leverage the expertise of a Guatemalan team member, who, supported by the cited work, asserts that the Spanish dialectal variation in MayanV extends beyond lexical differences and accurately represents the prevalent dialect among rural populations in the country.

5 Evaluating NMT Systems with MayanV

We develop bilingual and multilingual bidirectional NMT systems for Mayan languages and Spanish. Our aim is to assess the impact of the MayanV dataset on the translation quality of informal, familial domain texts. Thus, we consider baseline models that have not been exposed to this data during training and compare their performance to models trained with MayanV data. The validation and test sets consist exclusively of sentences from MayanV. Therefore, we will only evaluate language pairs for which MayanV data is available. Other languages still contribute to the training of multilingual models but are not explicitly evaluated. All corpora involved are parallel between Spanish and at least one Mayan language.

5.1 Experimental Settings

Bilingual and multilingual baseline models are trained using all relevant parallel corpora from OPUS, including Mozilla-I10n (Tiedemann, 2012) and bible-uedin (Christodouloupoulos and Steedman, 2015), as well as our own crawl of the Mayan versions of jw.org. Table 3 displays the data used to train the baseline models. To evaluate the impact of the MayanV dataset, we incorporate its data into the training pool (excluding dev and test sets), resulting in the setup shown in Table 4. We then train a new set of bilingual and multilingual models and compare them against the baseline models. Notice that, for some languages —e.g. Itza' (itz)—, we are unable to produce baseline models since the only data available are those taken from MayanV.

We use the same dev and test sets to, respectively, train and evaluate all models. We select 1 000 sentences from each corpus in MayanV as individual test sets, and 1 000 non-overlapping en-

ISO	jw.org			Mozilla I10-n			bible-uedin		
	Words (Mayan)	Words (es)	Sentences	Words (Mayan)	Words (es)	Sentences	Words (Mayan)	Words (es)	Sentences
cak	716 500	620 312	54 047	417 839	310 792	28 950	361 971	186 898	7 862
ctu	1 536 343	1 283 405	110 521						
ixl				17005	10 286	1 955			
kek	1 134 403	1 082 829	86 612				1 157 800	811 163	31 110
mam	1 711 960	1 523 974	124 051				265 748	185 460	7 799
poh	169 965	146 517	11 330						
quc	993 085	928 234	83 393	8 498	5 965	661	312 453	187 684	7 895
tzh	1 715 549	1 457 685	120 430						
tzo	3 238 511	2 942 428	234 599						
yua	3 554 344	3 452 737	263 500	4 361	2 440	306			

Table 3: Word and sentence distribution from the jw.org, Mozilla I10-n, and bible-uedin corpora.

ISO	Language	Words (mayan)	Words (es)	Sentences	train	dev	test
acr	Achi	6 994	7 657	1 343	–	343	1 000
agu	Awakatec	7 325	9 700	1 930	–	930	1 000
cac	Chuj	9 398	10 916	2 299	299 (299)	1 000	1 000
cak	Kaqchikel	1 496 310	1 118 002	90 859	90 859	–	–
ctu	Ch'ol	1 536 343	1 283 405	110 521	110 521	–	–
itz	Itza'	6 069	7 512	1 539	–	539	1 000
ixl	Ixil	27 893	22 423	4 280	2 280 (325)	1 000	1 000
kek	Q'eqchi'	2 310 937	1 915 942	121 883	119 883 (2 133)	1 000	1 000
kjb	Q'anjob'al	18 035	18 238	3 014	1 014 (1 014)	1 000	1 000
mam	Mam	1 727 413	1 543 091	134 943	132 943 (1 093)	1 000	1 000
poc	Poqomam	18 039	21 744	3 583	1 583 (1 583)	1 000	1 000
poh	Poqomchi'	176 444	153 666	13 117	11 117	787	1 000
quc	K'iche'	1 328 504	1 137 357	94 581	92 581 (632)	1 000	1 000
qum	Sipakapense	9 780	9 328	1 356	–	356	1 000
ttc	Tektitek	23 571	24 896	4 022	2 022	1 000	1 000
tzh	Tzeltal	1 818 858	1 586 344	140 276	138 276 (17 000)	1 000	1 000
tzj	Tz'utujil	12 283	11 404	2 519	519	1 000	1 000
tzo	Tzotzil	3 238 511	2 942 428	234 599	234 599	–	–
yua	Yucatec Mayan	3 558 705	3 455 177	263 806	263 806	–	–
Total		17 331 412	15 279 230	1 230 470	1 339 332	12 947	15 000

Table 4: Complete multilingual corpus. Parentheses indicate the amount of training sentences taken from MayanV.

tries as dev sets;[6] in cases where there are less than 2 000 entries —e.g. Achi (acr) or Sipakapense (qum)— we prioritise the test set and put the remaining sentences in dev set. After this, all remaining instances, if any, are included in the train set of the non-baseline models. Note that, as a result of the size of some corpora in MayanV, not all training sets have a MayanV corpus associated to them.

In addition to models trained from scratch, we examine (baseline and non-baseline) bilingual models resulting from the fine-tuning of the `nllb-200-distilled-600M` model implementation from the Huggingface Transformers library (Wolf et al., 2020). For this purpose, language tokens are added for each Mayan language, and NLLB-200's embedding layer is conveniently resized to accommodate them.

Following the methods described by Conneau and Lample (2019) to address data imbalance in multilingual models, we rebalance our combined training corpus by fixing the size of the largest language fraction by number of sentences, i.e. Yucatec Mayan, and upsampling with replacement all other fractions by computing

$$\lambda_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha},$$

where p_i is the number of sentences of the $i-$th language, N is the total number of languages in the corpus, and λ_i is the resulting number of sentences of the i-th language once increased. We empirically determine the optimal value of the exponent to be $\alpha = 0.7$.

We used the `fairseq` toolkit version 0.12 (Ott et al., 2019) to carry out our experiments, except for fine-tuning NLLB-200, for which we used Huggingface. We used byte-pair encoding (Sennrich et al., 2016) to tokenise our datasets into subword units, learning the joint vocabulary between Spanish and the combined multilingual corpora over 60 000 iterations. We used the Transformer model in its `base` configuration (Vaswani et al., 2017, Table 3),[7] with the added difference of using 8 000

[6] All dev sets are combined into a single dev set for multilingual models.

[7] Approximately 94M parameters.

warm-up steps and tied encoder-decoder embeddings. The multilingual models were trained on four parallel GPUs using mini-batches of 4 000 tokens, while the bilingual models were trained on a single GPU using mini-batches of 4 000 tokens as well. Validation was carried out every 5 000 updates, and the patience, based on the BLEU score on the dev set, was set to 20 validation cycles in order to ensure optimisation for all languages involved; each language pair was tested using the best-performing checkpoint as determined by validation. We applied label smoothing with a value of 0.1. The NLLB-200 model was fine-tuned on a single GPU, using mini-batches of size 8, and a max sequence length of 1 024 tokens. Validation was carried out every 500 steps, and the patience was set to 10 validation cycles. We used Adam for all training runs, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

5.2 Results and Discussion

Table 5 shows the BLEU (Papineni et al., 2002) scores[8] of all models trained from scratch, comparing baselines with those trained with MayanV, for the multilingual and bilingual translation systems; Table 6 shows the results of the corresponding baselines computed by fine-tuning the `nllb-200-distilled-600M` model. chrF2 scores, demonstrating similar trends, are presented in Appendix A.

The baseline models, which were trained over the available corpora that did not include any resources from MayanV, perform worse across the board when compared to models whose training data included MayanV. Within the bilingual runs, the NLLB-200 model fine-tuned to each individual language pair outperforms models trained from-scratch in almost all instances, with the notable exceptions being Tzeltal (tzh) and Q'eqchi' (kek). This increased performance is limited, however, since, as previously stated, NLLB-200 was not trained over any member of the Mayan language family, which limits the effects of positive transfer and curtails any other enhancement beyond the leverage of a few dozen loanwords taken from Spanish. Overall, while it is possible to trivially assert that these results reflect the similarity, or lack thereof, between the data in the train and dev sets and the data in the test set, we argue that these results also reflect the considerable discrepancy between (1) formal and archaic registers, such as the ones we expect to find in the Bible; (2) domain-specific content, such as the religious education and world news content of jw.org, and the more quotidian register found in MayanV, which is closer to the language Mayan speakers use in their day-to-day activities. These results also align with the dialectal divergences described in Section 4.2. The comparison between baseline results and those of models that include MayanV is also similar for multilingual models.

When comparing bilingual and multilingual models, the latter outperform the former in all but one instance: Ixil (ixl) to Spanish. For any given language with corpora other than MayanV, the results suggest the sizes of these corpora do not impact the performance of the translation task as much as we might expect; instead, the net size of the respective MayanV corpus seems to play a much greater role. For example, consider Mam (mam) and Tzeltal (tzh), whose jw.org corpora is similar in size, and, moreover, the former includes the New Testament in its training data; despite this, Tzeltal greatly outperforms Mam, seemingly because of the greater size of its MayanV corpus. Consider also Q'eqchi' (kek), with a sizeable jw.org corpus, the entirety of the Bible, and the second largest MayanV corpus; it, however, still compares to other languages with a smaller representation in the multilingual model and whose MayanV corpus is of similar size.

Since the data used in the bilingual models is a subset of the data used in the multilingual model, we conclude that the inclusion of other languages in the training run acts as favourable leverage in that we are able to benefit from the performance boost and positive transfer of multilingual models. In the case of Tzeltal, since it corresponds by far to the largest corpus in MayanV, we suggest that, by simply having more instances to observe, the model is able to generalise more broadly, even in the face of different registers. Overall, these results align with previous findings (Arivazhagan et al., 2019).

6 Conclusions

Very little work has been done with indigenous American languages regarding their informal, day-to-day use and interaction with the dominant languages of their surroundings, and our results re-

[8] Computed using `sacrebleu` (Post, 2018) with signatures `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1` and `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`

	maya-es				es-maya			
	bilingual		multilingual		bilingual		multilingual	
	Baseline	MayanV	Baseline	MayanV	Baseline	MayanV	Baseline	MayanV
acr	-	-	0.7	2.3	-	-	0	0.1
agu	-	0.0	0.3	1	-	0.0	0.1	0.3
cac	-	0.1	0.1	4.2	-	0.0	0	3
itz	-	-	0.6	0.9	-	-	0.1	0.1
ixl	0.3	8.4	0.2	5.8	0.0	1.3	0.1	4.3
kek	0.6	7.2	2.4	12.4	1.1	8.0	2.8	18.8
kjb	-	3.5	0.2	5.8	-	4.0	0	7.1
mam	0.9	3.8	1.1	6.5	0.2	1.0	0.4	2.9
poc	-	5.2	0.3	9.3	-	1.8	0	6.4
poh	-	0.2	1.3	5.5	-	0.3	2.7	4.8
quc	2.0	2.9	3	7.8	2.1	3.3	5.7	10.1
qum	-	-	0.6	1.6	-	-	0.1	0.1
ttc	-	6.9	0.7	10.5	-	7.1	0	12.2
tzh	1.8	68.1	1.8	70	1.0	58.0	2.3	64.1
tzj	-	0.0	0.2	4.3	-	0.1	0.1	3.2

Table 5: BLEU scores for the bilingual and multilingual Mayan–Spanish and Spanish–Mayan translation tasks over baselines and models trained from scratch with and without using MayanV.

	maya-es		es-maya	
	Baseline	MayanV	Baseline	MayanV
ixl	0.0	1.2	0.0	4.6
kek	1.1	9.7	2.7	5.4
mam	0.6	5.2	0.5	0.9
quc	5.2	5.7	4.3	5.4
tzh	3.5	51.1	2.4	18.4

Table 6: BLEU scores of the bilingual fine-tuned NLLB-200 model serving as baseline.

flect this unfortunate fact. We have developed and publicly released a curated set of parallel datasets between several Mayan languages and Spanish, which we call MayanV, focusing on the fact that the dialect and the register of the corpora is informal and non-domain-specific, which reflects a more common use case for the majority of native speakers. We train baseline bilingual and multilingual NMT Mayan-Spanish models from scratch, and fine-tune the NLLB-200 model for the bilingual case, and compare these with models whose train set is identical plus the addition of MayanV; we evaluate these models on a separate subset of MayanV and observe considerable improvements with respect to the baseline. This, along with a dialectometric analysis of the corpora involved, suggests that the vast majority of the available resources do not reflect the day-to-day usage of Mayan languages by their native speakers, nor do they facilitate the training and development of MT systems that might be useful for the most common use cases of the language. However, we do observe several instances of improvement in performance when comparing multilingual models with their bilingual counterparts, suggesting that this remains a valid pathway for developing and training ever-improving NMT models for Mayan languages.

Future work in the area of NMT of Mayan languages should focus on the mining and production of datasets that reflect a closer use case to that which is useful for rural and often times marginalised indigenous communities, who usually do not speak using overly formal or archaic language, nor have consistent access to the internet and its associated zeitgeist. Finally, interesting work is to be done by performing multilingual fine-tuning of larger pre-trained translation models, such as NLLB-200 itself, or MADLAD-400, whose purpose is to work with low-resource and endangered languages and which, sadly, minimise, or even outright exclude, Mayan languages in their current form.

Acknowledgments

Work funded by the Spanish Ministry of Science and Innovation, the Spanish Research Agency (AEI/10.13039/501100011033) and the European Regional Development Fund “A way to make Europe” through project PID2021-127999NB-I00.

Limitations

This paper has certain limitations, the primary being the relatively small size of the parallel corpora discussed therein. Despite significant efforts to construct this valuable resource, its limited scale may impact the generalisation and performance of any NMT systems developed for Mayan lan-

guages. The constrained amount of data available can potentially result in less robust models with limited capabilities to handle linguistic and contextual variability. Additionally, there is a risk that these systems could produce imprecise translations, posing potential safety and health concerns for users who rely on them. Despite these limitations, the obtained results provide a solid and valuable foundation for future research and efforts to expand resources available for Mayan languages.

Ethics Statement

The curated corpora and NMT models detailed in this paper actively advocate for the inclusion and promotion of indigenous languages. This aligns with the United Nations' goals, particularly during the International Decade of Indigenous Languages (2022-2032), emphasizing the crucial importance of preserving, revitalizing, and promoting indigenous languages globally. Our efforts will contribute to the development of language technology tools for Mayan languages, supporting linguistic diversity, and promoting equitable representation of language communities.

References

Academia de Lenguas Mayas de Guatemala. 2006. *Vocabulario awakateko*. Academia de Lenguas Mayas de Guatemala.

Academia de Lenguas Mayas de Guatemala. 2020. *Vocabulario itza'*. Academia de Lenguas Mayas de Guatemala.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Ryan Bennett. 2016. Mayan phonology. *Language and Linguistics Compass*, 10(10):469–514.

Ryan Bennett, Jessica Coon, and Robert Henderson. 2016. Introduction to mayan linguistics. *Language and Linguistics Compass*, 10(10):455–468.

Abraham Bote Tun. 2023. Urge crear "semilleros" de hablantes de maya desde educación básica: Indemaya. Retrieved on 2023-12-15.

Gerardo Caal Ixim, Rolando Choc Tzuy, Fidel Cac Culul, and Leonel Pacay Rax. 2004. *Vocabulario q'eqchi'*. Academia de Lenguas Mayas de Guatemala.

Lyle Campbell and Terrence Kaufman. 1985. Mayan linguistics: Where are we now? *Annual Review of Anthropology*, 14(1):187–198.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Michael D Coe and Mark Van Stone. 2016. *Reading the Maya glyphs*. Thames & Hudson.

José Luis Conguache Coj, Manuel Bernardo Malchic Nicolás, and Elvia Lucrecia García. 2001. *Diccionario bilingüe poqom-español*. Academia de Lenguas Mayas de Guatemala.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Jessica Coon. 2016. Mayan morphosyntax. *Language and Linguistics Compass*, 10(10):515–550.

Joseph Edmond DeChicchis II. 1989. *Q'eqchi'(Kekchi Mayan) variation in Guatemala and Belize*. Ph.D. thesis, University of Pennsylvania.

Gonzalo Donoso and David Sánchez. 2017. Dialectometric analysis of language variation in twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Association for Computational Linguistics*, pages 16–25.

David M Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World*, twenty-fourth edition edition. SIL International.

Nora C England. 2003. Mayan language revival and revitalization politics: Linguists and linguistic ideologies. *American anthropologist*, 105(4):733–743.

Mariajosé España. 2021. Plan de vacunación contra el covid ha excluido a comunidades indígenas. Retrieved on 2021-08-22.

Brian M Fagan. 1996. *The Oxford companion to archaeology*. Oxford University Press.

Brigittine M French. 2010. *Maya ethnolinguistic identity: violence, cultural rights, and modernity in Highland Guatemala*. University of Arizona Press.

Francisco García Mendoza, Gaspar Pérez Marcos, Juan Jacinto Lucas, Andrés Santizo Lucas, Pascual M. Domingo Pascual, Gaspar Miguel Pascual, and Lucy M. García Nicolás. 2003. *Vocabulario chuj*. Academia de Lenguas Mayas de Guatemala.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 Evaluation Benchmark for low-resource and multilingual Machine Translation. *arXiv preprint arXiv:2106.03193*.

Robert Henderson. 2016. Mayan semantics. *Language and Linguistics Compass*, 10(10):551–588.

Instituto Nacional de Estadística. 2018. Censo 2018. Resultados del Censo 2018.

Gaspar Ixcaya Ratzam, Pedro Culum Culum, Juan Quiacaín Navichoc, Mario Mendez González, and Antonio Baldomero Quiacáin González. 2019. *Vocabulario tz'utujil*. Academia de Lenguas Mayas de Guatemala.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2017. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32:167–189.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2021. NRC-CNRC machine translation systems for the 2021 AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 224–233, Online. Association for Computational Linguistics.

Philipp Koehn. 2020. *Neural Machine Translation*. Cambridge University Press.

Irina B. Kotenyatkina. 2019. Lexical peculiarities of the modern spanish language of guatemala. *RUDN Journal of Language Studies, Semiotics and Semantics*, 10:634–643.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier García, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *ArXiv*, abs/2309.04662.

SM Lakew, M Cettolo, and M Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 641–652.

Danny Law. 2014. *Language contact, inherited similarity and social difference: The story of linguistic interaction in the Maya lowlands*, volume 328. John Benjamins Publishing Company.

Manuel Laynez Ayay, Magdalena Guzmán Ceto, Diego Sambrano Rodríguez, Juan Caba Caba, Catarina Pérez Toma, María Asicona Canay, Pedro Cedillo Marcos, Miguel Chel Corio, Miguel Pérez de la Cruz, and Pablo de Paz. 2018. *Vocabulario ixil*. Academia de Lenguas Mayas de Guatemala.

Walter E Little. 2009. Language choice among mayan handicraft vendors in an international tourism marketplace. In *Imagining Globalization*, pages 85–105. Springer.

Egberto Catalino López Mejía, Germán Isaías Méndez de León, Jeremías Misael Pérez Hernández, Clemente Morales Berdúo, Cleotilde Vásquez Lucas, Pedro Alberto Velásquez Agustín, and Margarita Ordóñez Domingo. 2004. *Vocabulario mam*. Academia de Lenguas Mayas de Guatemala.

Pulkit Madaan and Fatiha Sadat. 2020. Multilingual neural machine translation in low resource settings.

Manuel Mager, Abteen Ebrahimi, Arturo Oncevay, Enora Rice, Shruti Rijhwani, Alexis Palmer, and Katharina Kann, editors. 2023. *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics, Toronto, Canada.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Laura Martinus and Jade Z Abbott. 2019. A focus on neural machine translation for african languages. *arXiv preprint arXiv:1906.05685*.

Juan Antonio Morán Mus, Irma Yolanda Cáal Có, Eliseo Chá Chá, Dilia Margarita Có Coy, Augusto Tul Raxche, Elvia Rosa Mó Cáal, and Felipe Ixim Jucub'. 2001. *Vocabulario poqomchi'*. Academia de Lenguas Mayas de Guatemala.

Lino Mauricio Méndez Pérez, Rosanio López López, Inocencio Pérez Simón, and Rodolfo Francisco Baltasar López. 2018. *Diccionario bilingüe tektiteko-español*. Academia de Lenguas Mayas de Guatemala.

Arturo Oncevay. 2021. Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201, Online. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, L. Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo KABENAMUALU, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan Van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, A. Oktem, Wole Akin, Ghollah Kioko,

Kevin Degila, H. Kamper, Bonaventure F. P. Dossou, Chris C. Emezue, Kelechi Ogueji, and A. Bashir. 2020. Masakhane - machine translation for africa. In *1st AfricaNLP Workshop Proceedings, 2020*, volume abs/2003.11529.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Andrés Pablo Escobar, Rosenda Antonia Pérez González, Daniel Pedro Mateo, Daniel López García, Santiago Juan Matías Lucas, Candelaria Pedro Juárez, and Carmelino Fernández González. 2003. *Vocabulario q'anjob'al*. Academia de Lenguas Mayas de Guatemala.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Doğruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open machine translation for low resource South American languages (AmericasNLP 2021 shared task contribution). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 218–223, Online. Association for Computational Linguistics.

Enrique Pato. 2023. Principales rasgos gramaticales del español de guatemala. *Zeitschrift für romanische Philologie*, 139(1):154–186.

Gilles Polian. 2017. *The Mayan Languages*, chapter Morphology. Routledge.

Gilles Polian. 2018. Diccionario multidialectal del tseltal tseltal-español. *Estudios e investigaciones. Ciudad de México: Centro de Investigaciones y Estudios Superiores en Antropología Social*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Robert Pugh, Francis Tyers, and Quetzil Castaeda. 2023. Developing finite-state language technology for maya. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 30–39.

Cristina Pérez Medrano and Patricia Delgado. 2010. *Vocabulario k'iche'*. Academia de Lenguas Mayas de Guatemala.

Michael Richards and Narciso Cojt'i Macario. 2003. *Atlas lingüístico de Guatemala*. Editorial Serviprensa Guatemala.

Sergio Romero. 2012. "they don't get speak our language right": Language standardization, power and migration among the q'eqchi'maya. *Journal of Linguistic Anthropology*, 22(2):E21–E41.

Sergio Romero. 2017. *The Labyrinth of Diversity: The sociolinguistics of Mayan languages*, chapter 15. Routledge.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Nllb Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Alejandro Teletor Velásquez, Elvy Esmeralda Sis Sis, Erito Tecú González, Otilia Ixcopal Cuxún, Lucía González Alvarado, Alberto Acetún Ramírez, and Juana Francisca Aj González. 2016. *Vocabulario achi*. Academia de Lenguas Mayas de Guatemala.

Juan Humberto Tema Bautista, Santos Serapio Ambrocio García, Otto Pérez Sales, and Werner Rubelsy Tema López. 2017. *Vocabulario sipakapense*. Academia de Lenguas Mayas de Guatemala.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Francis Tyers and Robert Henderson. 2021. A corpus of k'iche'annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis Tyers and Nick Howell. 2021. A survey of part-of-speech tagging approaches applied to k'iche'. In *Proceedings of the First Workshop on Natural*

Language Processing for Indigenous Languages of the Americas, pages 44–52.

Hiroto Ueda and Antonio Ruiz Tinoco. 2003. Varilex, variación léxica del espanol en el mundo: Proyecto internacional de investigación léxica. *Variación léxica del español en el mundo*, pages 141–278.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

A chrF2 scores

This appendix presents chrF2 scores (Popović, 2017) of our baseline models, including both from-scratch and NLLB-200 approaches, as well as bilingual and multilingual configurations. The results presented in Table 7 directly correspond to the BLEU scores detailed in Table 5, while Table 8 offers scores for the NLLB-200 models, aligning with the BLEU results in Table 6. The trends observed with both metrics are very similar, indeed aligning closely across all comparisons.

	maya-es		es-maya	
	Baseline	**MayanV**	**Baseline**	**MayanV**
ixl	9.3	21.7	5.0	9.3
kek	16.0	33.7	30.7	37.8
mam	23.2	27.6	12.7	24.7
quc	25.1	26.5	28.6	32.4
tzh	20.8	64.7	28.2	50.0

Table 8: chrF2 scores of the bilingual fine-tuned NLLB model serving as baseline, to be compared with those in Table 7.

	maya-es				es-maya			
	bilingual		multilingual		bilingual		multilingual	
	Baseline	**MayanV**	**Baseline**	**MayanV**	**Baseline**	**MayanV**	**Baseline**	**MayanV**
acr	-	-	14.3	18.4	-	-	10.7	12
agu	-	2.5	12.1	14.8	-	1.1	9.7	13.1
cac	-	2.1	10.6	21	-	1.9	10	27.2
itz	-	-	13.2	14.8	-	-	12.4	13
ixl	2.5	23.0	11.2	22.8	3.6	10.5	8.7	20.7
kek	18.3	27.4	18.1	32.6	19.9	30.5	28.2	42.9
kjb	-	22.0	13	25.3	-	21.9	11.1	29.3
mam	15.8	24.0	16.6	28.1	19.9	23.1	22.3	30.4
poc	-	22.4	13.2	28.5	-	16.7	10.4	32
poh	-	12.5	15.8	24.4	-	14.7	24.2	28.1
quc	19.9	20.8	20.7	27.3	24.2	25.1	30.9	35.7
qum	-	-	15.6	17.2	-	-	10.2	11.4
ttc	-	26.8	15	31.4	-	27.3	11	34.7
tzh	15.3	75.2	16.7	76.3	22.4	71.6	25.6	75.2
tzj	-	3.1	13.4	21	-	5.2	10.1	23.9

Table 7: chrF2 scores for the bilingual and multilingual Mayan–Spanish and Spanish–Mayan translation tasks over baselines and models trained from scratch with and without using MayanV.