FormosanMT: A Multilingual Parallel Corpus of the Formosan Language Family

Hunter Scheppat[1], Joshua Hartshorne[2], Sema Koc[1], Éric le Ferrand[3], and Emily Prud'hommeaux[1]
[1]Boston College, Chestnut Hill, Massachusetts, USA
[2]MGH Institute of Health Professions, Boston, Massachusetts, USA
[3]University at Buffalo, Buffalo, New York, USA
{scheppat,kocse,prudhome}@bc.edu, joshua.hartshorne@hey.com, ericlefe@buffalo.edu

Abstract

While the quality of machine translation (MT) between widely-spoken languages has improved dramatically in recent years, training robust MT systems for languages with fewer resources remains a challenge. Endangered languages, which often lack the speaker population and written tradition needed to create text resources, are at a particular disadvantage. Developing robust MT architectures for very low-resource settings is hampered by the lack of suitable parallel corpora. To address this challenge, we introduce FormosanMT, a set of MT-ready parallel corpora for the Formosan family of endangered languages indigenous to Taiwan. Together the corpora total nearly 500,000 Formosan-Mandarin and Formosan-English sentence pairs. We share scripts for extracting these corpora from public sources, along with customizable tools for filtering, normalizing, and partitioning the data. In addition, we provide a new tokenizer for Traditional Chinese writing compatible with the popular No Language Left Behind (NLLB) MT architecture, along with updated and improved code for fine-tuning NLLB for any low-resource language pair. Finally we distribute our fully trained NLLB and OpenNMT models for the Formosan languages to and from both Mandarin and English. In addition to serving as a valuable resource for the Formosan language speaker communities, our data, code, and models will be available to NLP researchers working on endangered and low-resource language MT.

Keywords: endangered language MT, endangered language resources

1. Introduction

Machine translation (MT) between languages with abundant training resources has made remarkable advances in recent years, particularly since the development of transformer models (Wang et al., 2022). Building robust models for under-resourced languages, however, remains challenging (Haddow et al., 2022). These challenges are amplified for the nearly 50% of the world's languages that are endangered, as these languages often lack a written tradition (Eberhard et al., 2024) and a large population of speakers able to produce data resources. This is particularly unfortunate given the potential utility of MT for supporting language documentation and preservation efforts (Zhang et al., 2020; Bird and Chiang, 2012).

Given the difficulty in generating parallel data for endangered languages – whether from scratch or from available translated educational, religious, or documentary texts – there are relatively few such corpora available for these languages. This lack of training data is a barrier to the development of MT training architectures and techniques designed to be effective specifically in extremely low resource settings (Agić and Vulić, 2019; Haddow et al., 2022).

In this paper, we take a step toward overcoming this obstacle with the creation of FormosanMT, a set of carefully prepared parallel corpora for the indigenous languages of Taiwan, along with scripts for reliably training MT models for two different architectures. The data, which is extracted from publicly available diverse data sources, consists of nearly 500,000 sentences pairing text in one of the 15 Formosan languages with translations in Mandarin or English. To ensure the continued growth of this resource, we include customizable scripts to extract, normalize, and filter the data, which can be used to recreate our corpora or to create new corpora as data is added to the public sources.

We additionally share a new tokenizer for Traditional Chinese writing compatible with the No Language Left Behind (NLLB) MT architecture (Costa-jussà et al., 2024), which currently supports only Simplified Chinese, along with improved and up-to-date code for fine-tuning NLLB to any new language pair, extending the utility of our work beyond Formosan. We successfully train and test MT models for all 15 languages with Mandarin and 4 of the languages with sufficient data with English using two popular MT architectures, the multilingual NLLB and the end-to-end OpenNMT architecture (Klein et al., 2020). We make these models available to serve as baselines for future work on MT for these languages.

This code and these corpora will serve as valuable resources not only for researchers and com-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3445–3455
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

munity members working to document and preserve the Formosan languages but also for anyone working on machine translation for endangered and under-resourced languages. In addition, because FormosanMT contains parallel corpora for 15 related languages, it provides a rich testbed for exploring the utility of incorporating related language data into an endangered language MT pipeline.

We summarize our contributions as follows.

- MT-ready parallel corpora in both English and Mandarin for the 15 Formosan languages
- Scripts for extracting, normalizing, and filtering the data for these sources
- A new NLLB-compliant tokenizer for Traditional Chinese writing
- New and improved code for fine-tuning NLLB
- Fully trained NLLB and OpenNMT models for all 15 Formosan languages
- All code and data freely available on GitHub.[1]

2. Related Work

Low-resource machine translation is a popular area of research, with a dedicated ACL workshop, LowResMT, now in its 9th year (Ojha et al., 2025). A variety of approaches have been proposed for low-resource MT, most of which rely on the transformer architecture (Vaswani et al., 2017), including mBART (Liu et al., 2020) and its extensions (Tang et al., 2020), MarianMT (Junczys-Dowmunt et al., 2018), mT5 (Xue et al., 2021), and No Language Left Behind (NLLB) (Costa-jussà et al., 2024). We leave discussion of pros and cons of these methods to future work, as the focus of our research here is resource creation, specifically for groups of endangered languages from the same family and region where limited speaker populations present challenges to the collection of large amounts of new text for training MT models.

A number of bespoke single-language MT corpora have been developed for endangered languages such as Cherokee (Zhang et al., 2020), Kotiria (Kann et al., 2022), Quechua (Ortega et al., 2020), Highland Puebla Nahuatl (Shi et al., 2021), and Ainu (Miyagawa, 2023), among many others. There is less work, however, on developing MT datasets containing multiple languages from the same family or region.

Some prior work on MT resource creation for groups of endangered languages has involved collecting new data in the target language, through transcription and translation of existing speech recordings, as in Bird and Chiang (2012)'s project to collect parallel data for 15 languages in Papua New Guinea. Most work, however, has relied on creating parallel corpora from existing organically parallel public sources, such digital dictionaries, educational materials, government publications, religious texts, and linguistic fieldwork archives. Such efforts include a corpus of four indigenous Columbian languages (Prieto et al., 2024); a dataset of four languages of Peru (Oncevay, 2021); parallel data from six mostly widely-spoken languages of Mexico (Martínez et al., 2020), and very small corpora for seven languages spoken in Eurasia (Mossolova and Smaïli, 2022). We follow this approach of compiling whatever resources are publicly available, with one exception: although the Bible is available for many of the Formosan languages, we do not include parallel corpora derived from the Bible in light of evidence that the significant domain and stylistic mismatch between this data and non-religious data can increase MT hallucinations (Domingues et al., 2024; Mayer and Cysouw, 2014).

There has been some prior work on MT specifically for the Formosan languages. Zheng et al. (2022) developed an mBART model to translate between Amis and Mandarin using a small corpus derived from the same ILRDF dictionary data that forms a portion of the data we include in the FormosanMT dataset. They followed this work with a project carrying out similar experiments, this time using an in-house transformer architecture, with all of the Formosan languages and Mandarin, again using only the ILRDF dictionaries. The focus of both studies was to investigate the utility of including lexical entries in the training pipeline and, in the second paper, to generate synthetic parallel data using the lexical entries. While related to our work, their results cannot be compared to ours since they use a subset of our data filtered and partitioned in its own way.

Our work extends our prior work (Scheppat et al., 2025), which focused on creating a shareable MT resource for Amis with two different translation languages and comparing multiple MT architectures across a carefully curated corpus spanning diverse public sources. In our new work, we distribute corpora not just for Amis but for all 15 Formosan languages, and we also share the code used to create (and recreate) this resource and to train models for two distinct MT architectures.

3. Data

3.1. The Formosan Languages

The Formosan language family of Taiwan is a subfamily of the larger Austronesian family, which includes languages like Tagalog, Maori, and Malay.

[1] https://github.com/FormosanBank/FormosanBank-MT

Language	ISO 639-3
Amis	ami
Atayal	tay
Bunun	bun
Kanakanavu	xnb
Kavalan	ckv
Paiwan	pwn
Puyuma	pyu
Rukai	dru
Saaroa	sxr
Saisiyat	xsy
Sakizaya	szy
Seediq	trv
Thao	ssf
Truku	trv
Tsou	tsu
Yami/Tao	tao

Table 1: Indigenous languages of Taiwan with their ISO 639-3 codes. Note that Seediq and Truku, though politically distinct languages, share the same ISO code.

Entirely unrelated to the Sino-Tibetan languages that are widely spoken today in Taiwan (e.g., Mandarin, Hokkien), all of the Formosan languages are endangered (Eberhard et al., 2024), with Kanakanavu and Thao classified moribund with only a handful of speakers still living. Table 1 lists the officially recognized indigenous languages of Taiwan – the Formosan languages along with the Austronesian language Yami – with their ISO 639-3 codes. While Seediq and Truku are recognized politically as distinct languages and appear separately in Table 1, they are linguistically classified as the same language with the same ISO code; as such, their available data is combined in our corpora. Yami does not belong to the Formosan language family, but as it is from the same broader language family and is indigenous to Taiwan, we include it in our work here.

Although the Formosan languages share certain features, such as complex morphology, an intricate system of voice, and a relatively small phonetic inventory, there is a large degree of linguistic variation across languages (Bellwood, 1984; Blust, 2019). This makes our corpus particularly suitable for exploring different methods of leveraging related language data during MT model training.

3.2. Data Sources

Much of the data used to create our parallel corpora comes from FormosanBank (Hartshorne et al., 2024; Mohamed et al., 2024)[2], an open repository of manually curated, transcribed, and translated Formosan speech and language data, compiled from various public sources with the permission of the organizations and communities that created the data. The following FormosanBank sources are included in all of the corpora:

1. **ILRDF Dictionary**: An electronic dictionary published by Taiwan's Indigenous Languages Research and Development Foundation (ILRDF) which contains dictionary entries with example sentences in Formosan and translations into Mandarin (Aboriginal Language Research and Development Foundation, 2023a).

2. **ePark**: A large educational website supported by the ILRDF. All texts are available in all the Formosan languages and Mandarin; many are also available in English and in recognized dialects of each Formosan language (Aboriginal Language Research and Development Foundation, 2023b).

3. **NTU Corpus**: Primarily fieldwork data consisting of conversations, stories, songs, and folktales in several of the Formosan languages, along with translations in Mandarin and English (Su et al., 2008; Sung et al., 2008).

4. **Presidential Apology**: An official apology issued by the president of Taiwan to the Indigenous people of Taiwan with translations in English and Mandarin.

In addition to these sources, which are available for all 15 languages, some very small language-specific sources have been cataloged in FormosanBank. Those sources, such as the Fey dictionary (Fey, 1986) with words and example sentences translated from Amis into Mandarin and English, are noted in the corpora and the scripts.

In addition to extracting pre-curated data from FormosanBank, we also extract the subtitles from the public videos produced by the ILRDF, which are available on the ILRDF website and YouTube.[3] The content primarily includes short-form, casual conversations with Formosan speakers with translations in Mandarin only. The videos typically range from 1 to 5 minutes in length. Amis and Paiwan are well represented, with relatively few videos available for the other languages. While these videos are fully public, their licensing restrictions are vague. We are working to gain explicit permission from the ILRDF to redistribute the subtitles in parallel corpus format; in the meantime, we provide scripts to pull the subtitles directly from the ILRDF website so that other users can extract the data exactly as we did when building our corpora.

We note that the Formosan languages are written using the Latin alphabet along with a few ad-

[2] https://ai4commsci.gitbook.io/formosanbank

[3] https://www.ilrdf.org.tw

ditional symbols, such as i and ^. While the Formosan languages were transcribed at various periods in Taiwan's history using other systems (e.g., katakana when the island was under Japanese control), the Latin alphabet has been used since the 17th century, and standardized orthographies using the Latin alphabet were adopted officially in 2005.

3.3. Data Normalization and Filtering

Detailed information about extraction of the parallel sentence pairs is described in the Appendix. As noted, the parallel corpora are drawn from two distinct sources, the FormosanBank repository and the ILRDF YouTube channel, and each has its own extraction scripts. The normalization and filtering process, however, is shared across the two sources.

Although the ePark and IRLDF dictionary resources were produced as written texts and prepared with the goal of serving as reference materials, the quality of the texts sourced from the ILRDF video transcripts and the NTU Corpus is more variable. The ILRDF videos contain spontaneous naturalistic language which can be difficult to transcribe and translate, while the NTU corpus consists largely of linguistic fieldwork transcribed and translated by linguists who often want to include commentary, questions, and notes in their translations. For these reasons, we take great care in normalizing and filtering the data.

After extraction, we normalize the text by applying Moses punctuation normalization when available, Unicode NFKC, removal of ASCII control characters, and removal of extra whitespace. We carry out the following normalization steps, in order:

1. Remove leading speaker tags that match a capital Latin letter followed by a colon (e.g.: "A:").
2. Delete short notes inside parentheses or brackets, length 1–10 characters, which provide commentary rather than translation.
3. Remove known artifact tokens that are Chinese strings translating to "full text record", "Chinese record", and "woman's full name".
4. Drop trailing commas (ASCII or Chinese).
5. Trim stray quotes/brackets.
6. Fix unusual spacing such as spaces before punctuation marks and spaces immediately after opening brackets/quotes and immediately before closing brackets/quotes.

The script then filters the data to remove sentence pairs where one or both sides contain text that suggests non-parallel content, including pairs where the:

- Mandarin side matches stage-direction phrases translating to "switch to the next item" or "I will stop here".
- Mandarin side is a page marker, bare enumeration, or year header.
- Mandarin side is simply repeated punctuation or sequences of more than two characters from a set of discourse particles (e.g., words corresponding to interjections like "ha" or "oh") identified by our Taiwanese collaborators.
- Formosan side contains Chinese characters.

Finally, we eliminate pairs where there is a large difference in the number of tokens or characters between the two sides of the pair, as follows. When both sides have two or more tokens, we compute the ratio, r, of target tokens to source tokens, and we discard pairs with $r \leq 0.2$ or $r > 8.0$. If either side has fewer than two tokens (typically because the pair is a dictionary entry), the ratio is instead calculated at the character level, and pairs are eliminated when $r \leq 0.05$ or $r > 20.0$. These thresholds were set through trial and error with the goal of minimizing the number of pairs discarded while still identifying particularly egregious non-parallel pairs.

We recognize that our filters are aggressive and might inadvertently eliminate valid sentence pairs, and we acknowledge the somewhat ad-hoc nature of some of our filtering heuristics. One reason that we are releasing not only the corpora but also our scripts for generating those corpora is that users can easily adjust and tune all of the filtering parameters and thresholds with command line flags to suit their particular needs and goals.

3.4. Data Partitioning

Creating unbiased data splits for a corpus derived from multiple diverse sources ranging from spontaneous narratives to individual dictionary entries requires a certain amount of caution and care. We take the following steps during data partitioning to minimize performance gains or losses that would simply be artifacts of the partitioning process:

- Duplicate sentence pairs are removed, leaving a single copy of that pair in order to prevent performance gains due to memorization.
- Sentence pairs where one side duplicates that side of an existing sentence pair are routed to the training partition.
- Sentence pairs where one side has been identified as a lexeme (dictionary entry) are routed to the training partition.

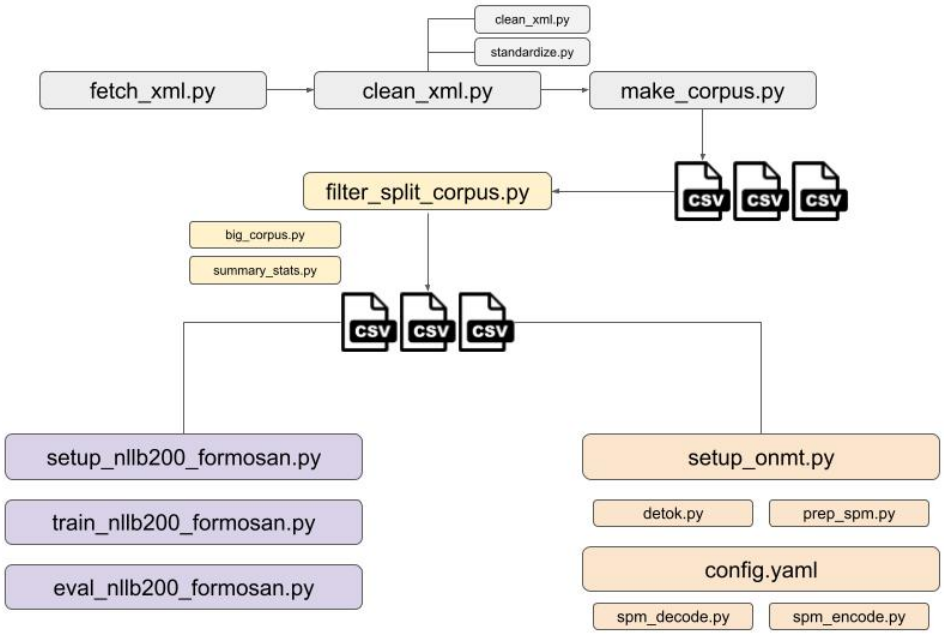

Figure 1: A flow chart detailing data extraction from FormosanBank, data cleaning, and model training.

Each individual data source is randomly partitioned with a fixed seed into an 80/10/10 train/dev/test partition. Partitioning each source separately facilitates more fine-grained evaluation of performance for different genres of text. As with the filtering parameters above, the random seed can be modified with a command line flag.

Table 2 presents the size of each parallel corpus after filtering and partitioning, organized by language pair. Aggregating across all Formosan languages yields 493,519 sentence pairs, of which 397,710 pair Formosan with Mandarin and 95,809 pairs Formosan with English. Dialect labels are occasionally provided in the source data. Although we do not consider dialect information in the partitioning or modeling in the work presented here, we keep any available dialect information in the source CSV files so that future work can use dialect information during partitioning, training, and testing. We also note that the corpora and the scripts themselves are versioned so that all counts reported in this section can be reproduced from the same sources without manual edits.

The full pipeline for data extraction, filtering, partitioning, along with model training is shown in Figure 1. As noted, additional information about the scripts used to extract the data is provided in the Appendix.

4. Method

In order to demonstrate the utility of the parallel corpora we have created, we train translation models using two distinct MT architectures: No Language Left Behind (NLLB) (Costa-jussà et al., 2024) and OpenNMT (Klein et al., 2020). These two neural MT architectures – the former a multilingual pretrained model that can be fine-tuned to a new language pair, and the latter an end-to-end approach – represent the two ends of the spectrum of available academic-quality neural MT architectures. While we could have explored additional architectures, we remind the reader that this is not an engineering paper but a resource paper with a focus on creating and describing a new large multilingual MT dataset for a family of endangered languages.[4]

4.1. OpenNMT

We first consider the end-to-end MT architecture, OpenNMT (Klein et al., 2020). Using OpenNMT-py[5] we train one bi-directional model per Formosan–English or Formosan–Mandarin pair. The system is a 2-layer LSTM encoder and decoder with input feeding and general global attention, `word_vec_size=256`, `rnn_size=512`, Adam at $1e-3$ with gradient clipping at 2, and token-based batching of 2048 tokens with `accum_count=2`, and dropout at 0.3. We construct the training corpus as a single concatenated bi-directional stream and control direction with a target tag prepended to each source line, indicating the language. To ensure a fair comparison with NLLB, we reuse the

[4]We note that our preliminary results using mBART were disappointing, as were our experiments with prompting GPT-5.

[5]https://opennmt.net/OpenNMT-py/main.html

Pair	Train	Valid	Test
ami ↔ eng	9,600	910	911
ami ↔ cmn	54,414	5,609	5,610
bun ↔ eng	9,351	857	857
bun ↔ cmn	33,627	3,095	3,097
xnb ↔ eng	3,667	365	367
xnb ↔ cmn	13,851	1,436	1,438
dru ↔ eng	11,810	978	980
dru ↔ cmn	35,913	2,988	2,990
pwn ↔ eng	6,342	566	567
pwn ↔ cmn	33,224	3,248	3,249
pyu ↔ eng	6,180	509	510
pyu ↔ cmn	22,715	1,906	1,907
ssf ↔ eng	1,812	136	138
ssf ↔ cmn	9,718	964	965
sxr ↔ eng	1,852	139	139
sxr ↔ cmn	8,299	773	774
szy ↔ eng	2,556	224	224
szy ↔ cmn	10,895	1,073	1,074
tao ↔ eng	1,914	154	155
tao ↔ cmn	10,176	1,029	1,030
tay ↔ eng	10,380	881	882
tay ↔ cmn	41,728	3,917	3,919
trv ↔ eng	7,268	667	668
trv ↔ cmn	29,170	2,862	2,862
tsu ↔ eng	2,499	222	223
tsu ↔ cmn	7,961	747	748
xnb ↔ eng	3,361	328	329
xnb ↔ cmn	12,347	1,197	1,198
xsy ↔ eng	2,808	261	262
xsy ↔ cmn	9,946	1,010	1,011
Total → cmn	333,984	31,854	31,872
Total → eng	81,400	7,197	7,212

Table 2: Corpus size, measured in number of sentence pairs, by language and partition.

same SentencePiece (Kudo and Richardson, 2018) model for subword segmentation and build OpenNMT vocabularies directly from the SPM-encoded files, which eliminates differing tokenization.

4.2. NLLB

For our NLLB models (Costa-jussà et al., 2024), we fine-tune the `facebook/nllb-200-distilled-600M`[6] sequence-to-sequence Transformer and train one bi-directional model between each Formosan language and Mandarin and between each of the four Formosan languages with sufficient English translations and English.

The model is controlled by language identifier tokens in the tokenizer vocabulary, so that at generation time we always set `tokenizer.src_lang` to the current source language code and force the target language by passing the `forced_bos_token_id` that corresponds to the desired target code. For compatibility across recent versions of the Transformers library, we additionally set `decoder_start_token_id` to the same id so that decoding deterministically starts in the correct target language. Labels never carry a language code; instead, we construct `decoder_input_ids` as the forced target is prepended to the target sequence. We append an explicit end of sentence (EOS) symbol to each label sequence and mask pad positions with `-100` so the loss ignores them. This design matches the behavior expected by NLLB while avoiding silent changes across library versions and ensures stable training in both directions with the same checkpoint.

4.2.1. Tokenizer Improvements

Our tokenizer and vocabulary adaptation proceeds by rebuilding the stock NLLB tokenizer to include all Formosan language codes and, for Traditional Chinese writing, improves segmentation when needed. The script first loads the original NLLB tokenizer and model, then reconstructs a fresh tokenizer instance that preserves all stock language codes and special tokens and appends new Formosan codes such as `ami_Latn`. We keep `<mask>` as the last additional special token.

We offer two paths to mitigate unknown tokens. In the character-addition mode, the script scans the corpus, identifies characters that would tokenize to `<unk>`, and adds only those characters whose frequency exceeds a user-set threshold. In the sentencepiece-merge mode the script trains a small SentencePiece (Kudo and Richardson, 2018) model on the dataset and merges its normal pieces into the NLLB SentencePiece model while ignoring any special pieces, which yields cleaner segmentation for Traditional Chinese writing (`zho_Hant`).

We use the sentencepiece-merge mode, and train a SentencePiece model on our dataset. After any tokenizer change, we resize the model's shared embedding matrix exactly once, warm start the new rows by averaging embeddings of the decomposition under the old tokenizer with a fallback to the unknown embedding, and seed newly added Formosan language-code rows that end with `_Latn` from `eng_Latn`. The script writes the updated tokenizer and model to disk together with a concise JSON report that records the new vocabulary size, the list of added codes, and the mapping from language codes to ids.

A small test generation section then verifies bi-directional generation for each new code in both directions `Formosan` to `zho_Hant` and `zho_Hant` to `Formosan` and, optionally, to `eng_Latn`. This makes the tokenizer surgery reproducible and evolves past previous fixes that are no longer compatible with newer versions of the `tokenizers` library.

[6] https://huggingface.co/facebook/nllb-200-distilled-600M

4.2.2. Training

Training uses a lightweight custom loop that samples direction per step with probability `p_src2tgt = 0.5`. At each step we set `tokenizer.src_lang` to the current source code and build batches on the fly, which avoids having to pre-build tokenized datasets and reduces memory pressure. Unless specified otherwise, our default hyperparameters are a batch size of 8, a maximum sequence length of 128 on both sides, and mixed precision on CUDA when available. We optimize with Adafactor to reduce memory footprint, using a constant schedule with 1,000 warmup steps, learning rate `1e-4`, weight decay `1e-3`, update clipping via `clip_threshold = 1.0`, and a global `max_grad_norm = 1.0`. Every `eval_interval` steps we run a tiny evaluation routine that computes average token loss in both directions on a held-out subset and prints short generations in both directions using a conservative decode configuration that sets both `forced_bos_token_id` and `decoder_start_token_id`, disables beam search for sanity checks, and uses `no_repeat_ngram_size = 3`, a small repetition penalty, and a mild length penalty to guard against degenerate loops. Checkpoints are saved periodically.

This approach is resilient to out-of-memory errors by clearing caches and continuing. All defaults can be overridden from the command line, including the language codes, column names in the CSV, sampling probability by direction, intervals, and decoding parameters for evaluation.

Our implementation attempts to update upon version-specific hacks found in the popular Medium tutorial for fine-tuning NLLB to new languages[7]. It also exposes tokenizer rebuilding as a single operation, and makes the language-code handling explicit. We avoid rewriting the `added_tokens.json` and related files by reconstructing the tokenizer with `additional_special_tokens` that already include the Formosan codes and by keeping `<mask>` last, which is compatible with recent `transformers` releases while remaining robust on earlier versions. We seed new language-code embeddings instead of leaving them random, we warm start newly added subword pieces, and we resize embeddings once after all tokenizer mutations, which prevents mismatches between tokenizer length and embedding rows. On the training side we never prepend language codes to labels, we always append EOS to labels, and we always set both `forced_bos_token_id` and `decoder_start_token_id` to the same target id, which makes the code stable under small changes in the generation API. We release both scripts so that others can reproduce the tokenizer augmentation and bi-directional fine-tuning pipeline for any low resource language.

[7] https://cointegrated.medium.com/how-to-fine-tune-a-nllb-200-model-for-translating-a-new-language-a37fc706b865

Direction	NLLB-200		OpenNMT	
	BLEU	chrF++	BLEU	chrF++
ami → cmn	**19.00**	**18.50**	13.90	13.10
cmn → ami	**10.07**	**34.16**	6.70	24.30
bun → cmn	**25.08**	**23.31**	19.70	18.50
cmn → bun	**7.85**	**35.65**	4.40	22.90
ckv → cmn	**24.33**	**22.62**	10.70	10.40
cmn → ckv	**25.88**	**48.88**	11.20	28.70
dru → cmn	19.97	20.09	**21.70**	**20.50**
cmn → dru	**5.77**	**29.74**	4.70	22.40
pwn → cmn	**16.54**	**17.03**	10.40	10.50
cmn → pwn	**8.63**	**35.45**	6.20	23.90
pyu → cmn	**24.22**	**23.76**	19.10	18.50
cmn → pyu	**14.10**	**39.11**	8.50	26.20
ssf → cmn	**25.09**	**23.27**	14.10	13.40
cmn → ssf	**19.80**	**47.28**	12.30	32.70
sxr → cmn	**14.38**	**15.43**	9.90	10.50
cmn → sxr	**7.10**	**39.67**	7.40	29.40
szy → cmn	**16.92**	**19.64**	10.60	11.80
cmn → szy	**20.20**	**43.40**	11.00	29.20
tao → cmn	**17.44**	**19.23**	9.10	10.10
cmn → tao	**18.59**	**40.02**	9.10	24.90
tay → cmn	**19.69**	**19.61**	16.40	16.70
cmn → tay	**5.89**	**26.99**	3.20	18.30
trv → cmn	**18.95**	**19.42**	18.00	17.00
cmn → trv	**10.38**	**31.21**	8.40	26.20
tsu → cmn	**15.16**	**17.26**	7.80	8.60
cmn → tsu	**11.11**	**34.27**	5.10	22.20
xnb → cmn	**26.90**	**25.94**	15.40	14.60
cmn → xnb	**21.31**	**51.79**	13.50	35.10
xsy → cmn	**20.79**	**21.06**	10.60	11.20
cmn → xsy	**22.45**	**45.54**	9.30	29.60

Table 3: Formosan ↔ Mandarin BLEU and chrF++ scores by direction with both architectures shown side-by-side. Boldfacing indicates the best performance per metric and language pair.

Direction	NLLB-200		OpenNMT	
	BLEU	chrF++	BLEU	chrF++
ami → eng	25.02	34.84	**33.50**	**40.60**
eng → ami	9.01	31.98	**10.20**	**32.00**
bun → eng	**43.08**	**53.55**	40.20	48.40
eng → bun	**10.08**	**38.11**	9.20	32.20
dru → eng	**29.84**	**40.95**	27.40	35.50
eng → dru	**4.66**	**28.11**	3.10	21.10
tay → eng	**41.02**	**49.19**	36.00	43.10
eng → tay	**4.19**	**23.88**	3.60	21.10

Table 4: Formosan ↔ English BLEU and chrF++ scores by direction with both architectures shown side-by-side. Boldfacing indicates the best performance per metric and language pair.

Amis	Reference	NLLB	OpenNMT
Mafiyok no fali ko kafong ako.	The wind blew off my hat.	My hat is off to the wind.	I ate that's wind.
Malalicalicay kita.	Let's all of us chat.	Let's have a meeting.	Let's dance
Matatodongay ko tayal nira.	His work is adequate.	His work is done right.	He has experienced over to work.
Ci Tipos ko ngangan no sava ako.	My younger sister's name is Tipus.	My elder sister's name is Tipos.	My dad's name is Api'.
A mimaan saw kiso i herek namicodad?	What are you going to do after school?	What will you be doing tomorrow afternoon?	What do you do after school?
Caay, a talaomah kita anini.	No, we are going to the fields.	No, we are going to the fields.	No, we are going to the fields.
Rakaten nira a tara i paisingan.	He walks to the hospital.	He took them to the fields.	He walks to the hospital.
Mangiliway kiso haw?	Have you lost weight?	Are you all right?	Do you have a motorcycle?

Table 5: Example output of NLLB and OpenNMT for Amis-English translation.

5. Results

For evaluation, we detokenize with the same SPM model used during training and compute both BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) using SACREBLEU (Post, 2018), enabling the built-in `zh` tokenizer for Mandarin targets. BLEU and chrF++ scores range from 0 to 100, with higher scores suggesting better translation quality. Table 3 present the evaluation results for translation to and from Mandarin for each Formosan language for each of the two MT architectures. Table 4 displays the same evaluation metrics for the four Formosan-English corpora with sufficient data to train a reasonable MT model.

A model was trained successfully for every language pair, even with the very small corpora for English, which attests to the quality and precision of our approach for normalizing and filtering the raw parallel data. There is a wide range of BLEU scores across languages, but with both architectures, translation into Formosan languages is generally much weaker than translation into either Mandarin or English. Interestingly, this holds both for NLLB, which as a pretrained multilingual model has prior knowledge of both English and Mandarin, and for OpenNMT, which does not. We also observe that even when BLEU scores are low when translating into a Formosan language, chrF++ scores are relatively high. Formosan languages have complex morphology; perhaps because chrF++ measures overlap at the character n-gram level, it allows for minor morphological differences between individual words in a hypothesis and the reference that would be penalized with BLEU.

We observe that translation into English yields much higher BLEU scores than translation into Mandarin or into any of the Formosan languages, even when using OpenNMT, which does not entail fine-tuning from an English-heavy multilingual model. While this may seem notable, it is likely an artifact of the limited domain and vocabulary of the Formosan-English datasets. Table 5 provides a few examples of NLLB and OpenNMT output, alongside the Amis input and a reference translation. We see that both models are prone to hallucination, but OpenNMT tends to yield more awkward and ungrammatical output.

Overall we see, perhaps unsurprisingly, that fine-tuning from a multilingual NLLB model generally outperforms the end-to-end OpenNMT LSTM approach. The one notable exception is translation between Amis and English, particularly from Amis to English, where OpenNMT is several points ahead of NLLB in both BLEU and chrF++. This was also observed in Scheppat et al. (2025), but the results here, which include models trained for 14 other languages, show that this seems to be particular to Amis. We note that the Amis corpus includes a number of additional small corpora from a variety of genres not found in the other languages, which may contribute to this effect.

6. Discussion and Conclusions

Machine translation has great potential as a tool for supporting endangered language documentation and reclamation, but the lack of parallel corpora for these kinds of languages presents obstacles to the development of new methods and architectures for truly low-resource MT. This paper describes the creation of FormosanMT, a large set of carefully normalized, filtered, and partitioned parallel corpora for the 15 indigenous languages of Taiwan, which can serve as a resource not only for the Formosan speaker communities but also for low-resource MT researchers. By sharing our scripts for creating these corpora, we ensure that as the data sources grow in size and diversity, our parallel corpora can grow in tandem. We provide new and modern code

for building MT models with one simple architecture and one state-of-the-art architecture that will allow others to replicate our findings and to train models for any pair of languages for which they have a corpus. Finally, we share the models we have trained, which can serve as baselines for future research in MT for the Formosan languages.

Our goals for future work focus on exploiting the multilingual and multigenre nature of FormosanMT. Some prior work shows that incorporating related language data during training can improve MT performance in low-resource settings. We plan to explore that approach via continued pretraining on the full FormosanMT corpus within NLLB and other architectures that involve fine-tuning from a pretrained multilingual corpus. Some of our corpora have more diversity across genres than others, and we would also like to explore improving source-level and genre-level performance with data augmentation techniques. Finally we plan to revive our preliminary work on LLM-based translation.

7. Limitations

One limitation of our work is that we trained baseline models using only two of the many available MT architectures that are compatible with small datasets. This is something we hope to do in future work, but we note that there is limited evidence suggesting that any one of the popular approaches is more successful than any other, including NLLB, for datasets of this size. We remain somewhat reluctant to spend resources training the large number of models required for each Formosan language for additional architectures.

A second limitation is that we did not consider dialect information in our partitioning, testing, or training. While the majority of our sources do not identify the dialect of the provided text, there may be non-trivial differences between dialects where they are identified, particularly at the lexical level.

8. Ethical Considerations

Respecting an Indigenous community's language sovereignty is crucial when working with endangered language data. The majority of our data was sourced from FormosanBank, whose creators have secured express permission from the ILRDF and other Indigenous groups and representatives to share and redistribute that data. The ILRDF videos are fully public but their licensing restrictions are not specified; for this reason we do not redistribute the associated texts but rather we share code that can allow other users to access the texts.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant #2319296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also thank Yuyang Liu, Li-May Sung, and the Indigenous Languages Research and Development Foundation, especially Akiw and Lowking Nowbucyang, for their contributions and for generously sharing their data.

9. Bibliographical References

Aboriginal Language Research and Development Foundation. 2023a. Online dictionary of aboriginal languages. https://e-dictionary.ilrdf.org.tw.

Aboriginal Language Research and Development Foundation. 2023b. Yuanzhumin yuyan leyuan (epark). https://web.klokah.tw/.

Željko Agić and Ivan Vulić. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *7th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics*, pages 3204–3210.

Peter Bellwood. 1984. A hypothesis for austronesian origins. *Asian Perspectives*, 26(1):107–117.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134.

Robert Blust. 2019. The austronesian homeland and dispersal. *Annual Review of Linguistics*, 5(1):417–434.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in ai tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages@ LREC-COLING 2024*, pages 283–293.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2024. *Ethnologue: languages of the world, 27th Edition*, volume 22. SIL International.

Virginia Fey. 1986. Amis dictionary. *Taipei: The Bible Society*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Joshua K. Hartshorne, Éric Le Ferrand, Li-May Sung, and Emily Prud'hommeaux. 2024. Formosanbank and why you should use it. In *Architectures and Mechanisms in Language Processing (AMLaP) Poster*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Katharina Kann, Abteen Ebrahimi, Kristine Stenzel, and Alexis Palmer. 2022. Machine translation between high-resource languages in a language documentation setting. In *Proceedings of 1st Workshop on NLP applications to field linguistics*, page 26.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. 2020. Cplm, a parallel corpus for mexican languages: Development and interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2947–2952.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*, pages 3158–3163.

So Miyagawa. 2023. Machine translation for highly low-resource language: A case study of ainu, a critically endangered indigenous language in northern japan. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 120–124.

Wael Mohamed, Éric Le Ferrand, Li-May Sung, Emily Prud'hommeaux, and Joshua Hartshorne. 2024. Formosanbank. https://ai4commsci.gitbook.io/formosanbank.

Anna Mossolova and Kamel Smaïli. 2022. The only chance to understand: machine translation of the severely endangered low-resource languages of eurasia. In *The Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT), COLING 2022*.

Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jonathan Washington, Nathaniel Oco, and Xiaobing Zhao, editors. 2025. *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*.

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.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second*

Conference on Machine Translation, pages 612–618.

Matt Post. 2018. A call for clarity in reporting {BLEU} scores. In *Proceedings of the Third Conference on Machine Translation*, page 186–191.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural preservation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 7–14.

Hunter Scheppat, Joshua Hartshorne, Dylan Leddy, Eric Le Ferrand, and Emily Prudhommeaux. 2025. Integrating diverse corpora for training an endangered language machine translation system. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 162–169.

Jiatong Shi, Jonathan D Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland puebla nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63.

Lily I Su, Li-May Sung, Shuping Huang, Fuhui Hsieh, and Zhemin Lin. 2008. Ntu corpus of formosan languages: A state-of-the-art report. *Corpus Linguistics & Linguistic Theory*, 4(2).

Li-May Sung, I Lily, Fuhui Hsieh, and Zhemin Lin. 2008. Developing an online corpus of formosan languages. *Taiwan Journal of Linguistics*, 6(2).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. Chren: Cherokee-english machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for amis-mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84.

Appendix: Data Extraction

Three scripts extract parallel sentence pairs from the FormosanBank repository. First, `fetch_xml.py` iterates through repositories in `formosanbank` and retains only documents whose root `TEXT` element carries `xml:lang` corresponding to the target Formosan language code. Second, `clean_xml.py` downloads the QC scripts from the FormosanBank project and applies them to the harvested XML. Third, `make_corpus.py` performs sentence extraction, iterating through sentence elements in the XML and pairing each source sentence with the first translation whose `xml:lang` is the desired target language. Each output row in the resulting CSV records the source sentence, the matched translation, the relative XML path, and the document-level `dialect` attribute, which we preserve unchanged for future researchers interested in exploring MT at the dialect level.

Acquiring the ILRDF video subtitles is more complex. `scrape_videos.py` identifies all available videos and collects each video's ID. For each video page, it requests the HTML, and iterates the per-sentence rows on the page; when a start-time attribute is present, it pairs the Formosan sentence with its Mandarin translation and preserves the start-time as an alignment anchor. The script writes one JSON file per video. Next, `make_xml.py` converts those JSON files into FormosanBank-style XML using `lxml.etree`. We then apply the process for converting FormosanBank XML into a CSV of parallel sentences.

We note that the structure and formatting of the FormosanBank and ILRDF YouTube resources are determined by their respective owners. As we are made aware of changes to these resources, we will update our extraction scripts to ensure compatibility.