Integrating diverse corpora for training an endangered language machine translation system

Hunter Scheppat[1], **Joshua K. Hartshorne**[2], **Dylan Leddy**[1],
Éric Le Ferrand[1], **Emily Prud'hommeaux**[1]
[1]Boston College, [2]MGH Institute of Health Professions
{scheppat,leferran,prudhome}@bc.edu,joshua.hartshorne@hey.com

Abstract

Machine translation (MT) can be a useful technology for language documentation and for promoting language use in endangered language communities. Few endangered languages, however, have an existing parallel corpus large enough to train a reasonable MT model. In this paper, we re-purpose a wide range of diverse data sources containing Amis, English, and Mandarin text to serve as parallel corpora for training MT systems for Amis, one of the Indigenous languages of Taiwan. To supplement the small amount of Amis-English data, we produce synthetic Amis-English data by using a high quality MT system to generate English translations for the Mandarin side of the Amis-Mandarin corpus. Using two popular neural MT systems, OpenNMT and NLLB, we train models to translate between English and Amis, and Mandarin and Amis. We find that including synthetic data is helpful only when translating to English. In addition, we observe that neither MT architecture is consistently superior to other and that performance seems to vary according to the direction of translation and the amount of data used. These results indicate that MT is possible for an under-resourced language even without a formally prepared parallel corpus, but multiple training methods should be explored to produce optimal results.

1 Introduction

The potential of language technology to support endangered language documentation and revitalization efforts is well established though not always effectively realized (van Esch et al., 2019). Machine translation (MT) in particular has been cited as a useful tool (Zhang et al., 2020; Bird and Chiang, 2012). First, translation from an indigenous language into a more widely spoken language is a common, if not required, part of generating linguistic documentation. This also ensures that understanding of the language will continue even if the language ceases to be used regularly (Bird and Chiang, 2012). Second, MT is often proposed as a way to make languages more accessible to language learners in Indigenous communities where younger generations were not raised speaking the language(Pinhanez et al., 2024). Finally, MT is appealing to NLP researchers because generating a new dataset only requires the expertise of a speaker to produce a translation; translation does not require complex software to control audio playback or alignment of audio with transcription, as speech transcription might, or extensive annotator training as part-of-speech tagging or parsing would.

Unfortunately, building a reasonable MT system with the quantity of parallel data typically available for an endangered language is remarkably challenging. There are few existing parallel corpora, and since nearly half of the world's languages lack an established writing system or written tradition (Eberhard et al., 2024), there are generally very few texts in the target language that can be translated in order to create a parallel corpus.

In this paper, we describe a broad effort to compile two parallel corpora – one with English and one with Mandarin – for Amis, one of the 16 recognized indigenous languages of Taiwan. We use nine different public sources[1] for our parallel data, which range from digital dictionaries to pedagogical materials to websites with user-contributed translations to YouTube videos curated and translated by Taiwan's Indigenous Language Research and Development Foundation. Since very little of this data includes English translations, we also generate synthetic Amis-English parallel data by using Mandarin as a pivot language, using high-quality MT to produce English translations of the Mandarin side of the Amis-Mandarin parallel data.

Using two different popular neural MT archi-

[1] Please see the Ethical Considerations section for details about our data use agreements.

tectures – the end-to-end OpenNMT framework (Klein et al., 2020) and the No Language Left Behind architecture for fine-tuning from a large pretrained multilingual model (Costa-jussà et al., 2024) – we build models to translate between Amis and Mandarin, and Amis and English. Even with our small and diverse "found" datasets, we are able to achieve reasonable BLEU and chrf++ scores. Supplementing the Amis-English parallel corpus with pivot data yields improvements when translating to English but not to Amis. Interestingly, we find that neither architecture consistently outperforms the other. The results suggest that compiling parallel data from diverse sources can create a corpus sufficient for training reasonable MT models. The interactions between architecture, corpus size, and translation direction, however, require additional study.

2 Background

2.1 Amis language

Though spoken in Taiwan, Amis (ISO 639-3 language code ami) is unrelated to Mandarin or other Sinitic languages. Rather, it is a member of the Formosan branch of the Austronesian language family, one of the largest families in the world both by number, with around 1,200 extant languages, and by geographic spread, ranging from Malagasy in the West to Rapanui in the East, and from New Zealand in the South to Taiwan in the North. Amis, the most widely spoken of the 16 Formosan languages with just over 100,000 speakers, has five officially-recognized dialects and is classified by Ethnologue as *threatened* (Eberhard et al., 2024).

Amis, like other Formosan languages, has a number of typologically unusual features (Li et al., 2024). It has primarily VSO word order. It has a limited phonetic inventory, with a only three vowels. It makes use of reduplication as part of the grammar, and its lexical roots are not easily categorized by part of speech. Most famously, Amis and other Formosan languages have a complex grammatical voice system. In short, Amis bears little resemblance to the languages used to train most multilingual models, including the multilingual NLLB model.

2.2 Endangered language MT

As noted previously, MT is recognized as a potentially useful tool for language documentation (Bird and Chiang, 2012; van Esch et al., 2019). A number of MT systems for endangered and indigenous languages have been developed for research or demonstration purposes, including (among many others) Cherokee-English (Zhang et al., 2020), Kotiria (Kann et al., 2022), Quechua (Ortega et al., 2020), Highland Puebla Nahuatl, and Ainu (Miyagawa, 2023).

There is a small amount of prior work on MT for Amis specifically. Zheng et al. (2022) created a small parallel Amis-Mandarin corpus using a subset of the ILRDF data that we use and an associated dictionary (see Section 3.1). Using an mBART-based transformer model, they trained models to translate between Amis and Mandarin. In a follow-up paper, Zheng et al. (2024) continued this work with multiple Formosan languages exploring the impact of including additional data, particularly dictionaries and lexica, as well as synthetic data. While their results are not directly comparable to our Amis-Mandarin results given the very different training corpora, their results without data augmentation are comparable to those we present here. We note, however, that we consider translation both to Mandarin and to English. In addition, unlike our work, this prior work does not compare the performance of different MT architectures.

3 Data

3.1 Data sources

The following data sources were used to create parallel corpora for our experiments translating between Amis and English and between Amis and Mandarin.

1. **ILRDF Videos**: Videos and manually-produced captions created by the Indigenous Language Research and Development Foundation (ILRDF) of Taiwan. The content primarily includes short-form, casual conversations with Amis speakers, with translations provided in Mandarin. The videos typically range from 1 to 5 minutes in length.

2. **Presidential Apology**: An official apology issued by the president of Taiwan to the Indigenous people of Taiwan. Although brief, the document contains high-quality text with long sentences, available in Amis, English, and Mandarin.[2]

[2] https://www.president.gov.tw/NEWS/20603

3. **Bible**: A short, user-generated and unverified section of the New Testament in Amis with English translations.

4. **ePark** (Aboriginal Language Research and Development Foundation, 2023b): A large electronic education website supported by ILRDF. All texts are available in Amis and Mandarin; many are also available in English and one or more Amis dialects.

5. **Glosbe**: An online community-developed dictionary similar to Wikipedia, which includes user-contributed Mandarin translations and definitions.[3]

6. **ILRDF Dictionary** (Aboriginal Language Research and Development Foundation, 2023a): An electronic dictionary published by ILRDF which contains extensive example sentences in Amis and translations into Mandarin.

7. **Zheng Corpus**: Zheng et al. (2022) made available a dataset based on the above ILRDF Dictionary resource that contains some new text, with translations into Mandarin.

8. **NTU Corpus** (Su et al., 2008): 16 narratives and 2 conversations in Amis, with free translations in English and Mandarin.

9. **Fey Dictionary** (Fey, 1986): Amis dictionary compiled by Virginia Fey, with example sentences translated into Mandarin and English.

3.2 Data acquisition and alignment

The Glosbe, Bible, and ILRDF Video texts were downloaded from the Web, and the Presidential Apologies were extracted from PDFs. The ILRDF Dictionary texts were obtained through the ILRDF API. The NTU Corpus and ePark were provided to the authors by the owners.

The ILRDF video data posed several challenges. First, many instances of code-switching occurred, where Amis speakers switched to Mandarin mid-sentence, resulting in Chinese characters appearing within the Amis text. Some sentences contained only Mandarin, as speakers switched languages for extended periods. Additionally, the quality of the translation pairs was sometimes inconsistent. For instance, many pairs included the Mandarin word for “unknown”, indicating that the translator was unsure of the Amis speaker’s meaning. Furthermore, the translations often included descriptions of non-verbal sounds, such as “leaves rustling” which did not appear in the original Amis text, which we attempted to automatically filter out.

The quality of Glosbe data was also challenging. The text often included unusual formatting, and there was significant overlap between Glosbe and other texts, such as the Bible. The Glosbe website was scraped by searching for common words in the Amis language, as direct access to all sentences for one language was unavailable. After the search, duplicates were removed, and the remaining sentences were formatted into parallel text.

The text of the Presidential Apology was manually aligned separately for each translation pair. As a result, the sentence counts differ in the two parallel corpora.

The NTU Corpus was prepared by linguists specializing in Amis and fully bilingual in English and Mandarin, yielding reliable and high quality translations. Similarly, because the ePark corpus consists of officially-produced and verified language educational material, it is of very high quality. We note that most of the ePark texts, rather than being produced initially in Amis, were instead translated *to* Amis from Mandarin or English.

The Bible data used here – a user-generated subset of Bible verses – was available online prealigned. We note that the English side of this corpus features the archaic language found in the King James version of the Bible, which may render this corpus less useful for translation to contemporary English.

The Fey dictionary included translations of sentences, short sentence fragments, and individual words. Occasionally, a single Amis sentence had multiple valid English translations. In these cases, we included each English translation as a distinct pair with the original Amis sentence.

3.3 Data preprocessing

Data processing focused primarily on ensuring reliable alignments and translations and correct formatting. To effectively address the issue of noisy translation pairs – those that either do not accurately reflect the source content – we implemented a fertility heuristic. This heuristic was designed to filter out sentence pairs exhibiting significant discrepancies in length between the source and target texts. The assumption underlying this approach is

[3] https://glosbe.com/

Corpus	Language	Sentence Pairs (w/ pivot)	Tokens (w/ pivot)	Types (w/ pivot)
ILRDF Videos	eng	(38,780)	(268,006)	(10,149)
	ami	38,780	227,074	21,469
Presidential Apology	eng	92	1,559	532
	ami	92	1,573	422
Bible	eng	512	11,676	1,482
	ami	512	11,469	1,679
ePark	eng	21,699 (27,904)	87,693 (143,319)	5,074 (8,138)
	ami	27,904	127,328	13,298
Glosbe	eng	(1,305)	(18,732)	(2,759)
	ami	1,305	18,340	2,752
ILRDF Dictionary	eng	(17,763)	(99,639)	(7,100)
	ami	17,763	80,012	8,756
NTU Corpus	eng	922	8,881	1,252
	ami	922	8,881	1,595
Fey Dictionary	eng	2,180	11,827	2,248
	ami	2,180	9,621	2,273

Table 1: Corpus sentence pair, token, and type counts for the English-Amis corpora. For English, counts without pivot data appear outside parentheses, while counts including pivot data appear inside parentheses. Token counts are all word-based. Note that the pivot data includes only a subset of the full Mandarin-Amis dataset.

that a large difference in length could indicate a potential misalignment or a translation that deviates considerably from the source material. We also implemented hard-coded detection mechanisms to identify noisy or incorrect translation pairs. For instance, in many cases in the ILRDF Video data where the translator failed to understand the original speech, the translation was labeled as “indistinct” or “no Chinese record”. Such sentence pairs were removed from the corpus. Further processing was centered on preparing data for machine translation. We utilized the Moses library to perform spellchecking and to harmonize punctuation.

Due to the overlap between sources and the inclusion of multiple dialects of Amis in some of the sources, the corpus contained both duplicate translations and many-to-one translation mappings, where a single word or phrase was associated with multiple possible translations in the other language. Duplicate translation pairs were retained but exclusively allocated to the training data. This approach ensures that no sentence pairs appears in both the training and testing sets.

We note that we did not attempt to distinguish among the five dialects represented in the corpora. Given the very limited amount of data for training, we treated all Amis data as one language. We plan to address the complexities of dialectal variation in our future work.

3.4 Pivot data creation

While all of the Amis words, phrases, and sentences in the datasets had Mandarin translations, only a small percentage had English translations. To augment the much smaller Amis-English corpus, we used Mandarin as a pivot language to create new Amis-English pairs by translating the Mandarin side of the Amis-Mandarin pairs into English. For this task, we used the DeepL API[4], which offers a free tier of 1,000,000 characters per month. The Mandarin text from each corpus was submitted to the DeepL API in batches, specifying English as the target language and Mandarin as the source language. This process increased the size of the English-Amis corpus from 25,405 pairs to 89,458 pairs. While this was an efficient way to synthesize new training data, we did observe that the translations did not always faithfully render the general style or tone of the original Mandarin text.

[4] https://www.deepl.com/en/pro-api

Corpus	Language	Sentence Pairs (w/ pivot)	Tokens (w/ pivot)	Types (w/ pivot)
ILRDF Videos	ami	41,459	241,295	24,354
	cmn	41,459	395,066	3,303
Presidential Apology	ami	33	1,929	530
	cmn	33	3,536	580
ePark	ami	48,071	319,138	19,397
	cmn	48,071	543,948	3,039
Glosbe Amis	ami	5,860	91,201	4,242
	cmn	5,860	160,315	1,748
ILRDF Dictionary	ami	5,482	37,140	8,054
	cmn	5,482	61,383	2,462
Zheng Corpus	ami	15,022	48,764	11,994
	cmn	15,022	93,734	2,822
NTU Corpus	ami	742	7,718	1,282
	cmn	742	11,650	904
Fey Dictionary	ami	2,478	10,619	2,436
	cmn	2,478	17,706	1,924

Table 2: Corpus sentence pair, token, and type counts for the Mandarin-Amis corpora. Token counts for Mandarin are based on subword-unit token counts from the NLLB tokenizer, while token counts for Amis are word-based.

Architecture	NLLB		OpenNMT	
Eval Metric	**BLEU**	**chrf++**	**BLEU**	**chrf++**
Amis -> English without pivot data	11.35	25.50	14.68	28.14
Amis -> English with pivot data	14.55	29.59	20.44	32.74
English -> Amis without pivot data	10.78	37.13	N/A	N/A
English -> Amis with pivot data	10.38	37.10	8.34	31.87
Mandarin -> Amis	15.15	36.25	17.10	36.70
Amis -> Mandarin	23.83	26.64	28.10	31.70

Table 3: MT output evaluation across architectures (NLLB vs. OpenNMT) and training corpora (Amis-English with and without pivot data, Amis-Mandarin). N/A indicates that the model was apparently unable to learn, yielding output consisting entirely of <unk> tokens.

3.5 Data partitioning

The Amis-English corpus contained 25,405 without the pivot data and 89,458 sentence pairs including the pivot data. The Amis-Mandarin corpus contained 119,147 sentence pairs. We partitioned the datasets as follows. First all duplicate pairs were removed from each corpus. From the remaining sentences pairs, approximately 5% of the pairs from each corpus were selected to form the test set for that corpus. Duplicate sentences were added back to the corpus, and the remaining sentences pairs of each corpus formed the training data.

4 Method

While there are various approaches to MT in extreme low-resource settings, we focus on two popular approaches: an older but reliable end-to-end sequence-based MT architecture, OpenNTM (Klein et al., 2020), and fine-tuning with the multilingual No Language Left Behind (NLLB) architecture (Costa-jussà et al., 2024). For the OpenNMT training, we followed the most recent version of the OpenNMT tutorial.[5]. For NLLB, our starting

[5] https://github.com/ymoslem/OpenNMT-Tutorial

point was a notebook[6] originally used to fine-tune the NLLB-200 distilled 600M model to translate between the Turkic language Tyvan and Russian.

When using this NLLB framework, we initialized the tokenizer for Amis with a base configuration set for the only available related language, Tagalog, noted internally as `tgl_Latn`. We additionally added a new language token specific to Amis, identified within our system as `amis_Latn`. This required modifying the tokenizer's vocabulary to include Amis and adjusting the model's embeddings to accommodate this addition. The embedding for the new Amis token was initialized using the embeddings of the Tagalog language, leveraging the linguistic similarities to enhance the model's performance without extensive retraining from scratch. This process also involved repositioning certain tokenizer elements, such as the mask token, to maintain the tokenizer's integrity and functionality after the introduction of new language components.

Mandarin in Taiwan is written using traditional Chinese orthography. The NLLB tokenizer documentation indicates that the `zho_Hant` tag can accommodate both simplified Chinese and traditional Chinese orthography, but we found that only 60% of the traditional characters in our data were accounted for by the tokenizer. We addressed this problem by training a custom SentencePiece (Kudo and Richardson, 2018) tokenizer on our Mandarin data and then inserting the resulting missing tokens into the token set for the NLLB `zho_Hant` tokenizer.

As NLLB models are inherently bidirectional, we built three models: Amis-English with no pivot data; Amis-English with pivot data; and Amis-Mandarin. Within OpenNMT, whose basic design is for unidirectional training, we trained six models: Amis->English and English->Amis without pivot data; Amis->English and English->Amis with pivot data; Amis->Mandarin; and Mandarin->Amis. We evaluate the output of these models on our test data using two metrics: BLEU, the long-standing MT evaluation metric based on word n-gram precision, and chrf++ ("CHaRacter-level F-score"), which uses character-based, rather than word-based, n-grams.

5 Results

Table 3 presents the BLEU and chrf++ scores for each model trained, in each direction, for each of the two MT architectures. We first consider the impact of including synthetic pivot data. We see that when translating from Amis to English, the inclusion of pivot data increases performance as measured by both metrics – at times rather dramatically – under both MT architectures. Strangely, including pivot data when translating to Amis does not yield improvements. With OpenNMT, the very small amount of unpivoted Amis-English data was insufficient to train a model. The addition of pivot data facilitates the production of actual output, but BLEU scores are weak.

We now turn to a comparison of NLLB with OpenNMT. When translating from Amis to English, OpenNMT outperforms NLLB by several points in terms of both BLEU and chrf++. When translating to Amis from English, NLLB outperforms OpenNMT, which fails to yield output at all for the condition with no pivot data. For the Amis-Mandarin models, OpenNMT again holds a small advantage over NLLB, with higher BLEU and chrf++ scores in both directions. Interestingly, translation to Amis using NLLB, whether from English or Mandarin, yields very high chrf++ scores. We conclude that both architectures show promise for translation in these low-resource settings, with each architecture outperforming the other under certain conditions.

Table 4 shows a few example outputs for OpenNMT and NLLB trained on the larger Amis-English dataset that included pivot data. Recall that OpenNMT yields higher BLEU and chrf++ scores than NLLB when translating to English. In all cases, the translations are reasonable. We see that OpenNMT appears more likely to produce verbatim matches for the reference or text with more shared unigrams. We also observe that the NLLB output sometimes includes words found in the Amis input, something that almost never happens in the OpenNMT output. Overall, while the BLEU and chrf++ scores reported for this translation direction are higher for OpenNMT, the NLLB translations often capture the gist of the reference without necessarily generating a verbatim match, which will negatively impact BLEU scores. The final example is typical of output produced from input sentences drawn from the Bible. These consistently yielded Biblical language unrelated to the

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

Amis	Reference	OpenNMT	NLLB
Mataturuturud kita tu tayal	Let us pass on the work from one to the other	one by one of our work is continued	let us work together
Maulah kaku aci kaka ^aku a rumadiw	Both my brother and I like to sing	both my brother and i like to sing	i also like to sing with my brother and sister
Aciyah adihayay tu ku tayal isu	Ouch you have got a lot on your plate	wow you guys are a lot of work	wow i have got a lot of work
Adihay ku heci nu kilang nira	His fruits were plenty	he has many fruit	the fruits of its trees are numerous
Ahecid ku nanum nu liyal	The sea is salty	the sea is salty	the sea water is salty
Hay fangcal ku keru ^nu maku	Yes i dance well	yes i dance well	yes my dances are great
Ira ku rengus I umaumahan	There is grass in the field	is there any grass here	rengus is in the field
Matngiltu namu ku nanu sapi'met a limuut nu itiyaayhu a tamdaw tuya aka pipatay tu tamdaw o mipatayay a tamdaw i u mamasawkit sanay	You have heard that it was said to the people long ago, 'Do not murder, and anyone who murders will be subject to judgment.'	amen i say to you whatsoever you shall bind upon earth shall form and shall cast on enemy	and when jesus was come into the house of the ruler and saw the minstrels and the multitude making a rout

Table 4: Example Amis-English sentence pairs, along with the predicted output of OpenNMT and NLLB when trained on data including the pivot data.

actual content of the reference sentence for both architectures.

6 Conclusions

While accurate machine translation offers utility for supporting language documentation, training a robust model requires a large parallel corpus that would be difficult to acquire for most endangered and indigenous languages. In this paper we mined a wide variety of diverse existing corpora containing parallel data in order to produce MT-ready corpora for Amis-English and Amis-Mandarin translation. We were able to achieve promising BLEU and chrf++ scores under some conditions, but many questions remain about the utility of the pivot data for translation into Amis and about the performance contrasts between NLLB and OpenNMT.

In our future work, we plan to carry out data ablation studies to determine the individual contributions of each of the component corpora. In particular we suspect that the Bible data may not be appropriate given the style and content of the other corpora. Given the ethically and practically dubious value of Bible data in low-resource MT and other NLP tasks (Liu et al., 2021; Domingues et al., 2024), we may see improvements while recognizing the potentially harmful effects of using culturally irrelevant texts. We also would like to incorporate dictionary entries in a more effective way following (Zheng et al., 2024), who showed success across a large number of languages in the Formosan family.

Ethical considerations

When working with an indigenous language, it is necessary to ensure that the community to whom the language belongs is a willing and active participant in the research. While the data used in our project is freely available for download, we have taken extra steps to gain the explicit permission of the Indigenous Language Research and Development Foundation (ILRDF) and the managers of the ePark indigenous educational organization to use their Amis data. All of our models will be shared with these organizations and the Amis community.

Acknowledgments

The authors thank Yuyang Liu, Li-May Sung, and the Indigenous Languages Research and Development Foundation, especially Akiw, Lowking Nowbucyang, and Yuyang Liu, for generously providing data. 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.

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/.

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

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 Under-resourced 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*.

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. *Preprint*, arXiv:1808.06226.

Paul Jen-kuei Li, Elizabeth Zeitoun, and Rik De Busser, editors. 2024. *Handbook of Formosan Languages: The Indigenous Languages of Taiwan*. Brill's Handbooks in Linguistics. Brill.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 44–50.

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.

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.

Claudio Pinhanez, Paulo Cavalin, Luciana Storto, Thomas Fimbow, Alexander Cobbinah, Julio Nogima, Marisa Vasconcelos, Pedro Domingues, Priscila de Souza Mizukami, Nicole Grell, et al. 2024. Harnessing the power of artificial intelligence to vitalize endangered indigenous languages: Technologies and experiences. *arXiv preprint arXiv:2407.12620*.

Lily I-wen 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 and Linguistic Theory*, 4(2):291–294.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22.

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.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11248–11259.