A Parallel Corpus and Dictionary for Amis-Mandarin Translation

Francis Zheng, Edison Marrese-Taylor, Yutaka Matsuo
Graduate School of Engineering
The University of Tokyo
{francis, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

Amis is an endangered language indigenous to Taiwan with limited data available for computational processing. We thus present an Amis-Mandarin dataset containing a parallel corpus of 5,751 Amis and Mandarin sentences and a dictionary of 7,800 Amis words and phrases with their definitions in Mandarin. Using our dataset, we also established a baseline for machine translation between Amis and Mandarin in both directions. Our dataset can be found at https://github.com/francisdzheng/amis-mandarin.

1 Introduction

Amis is a minority language spoken on the east coast of Taiwan and has been described as a vulnerable or endangered language (Moseley, 2010; Edmondson et al., 2005; Kuo, 2015; Liu, 2011). Though there have been some efforts to preserve the language through education and linguistic research (Wu and Lau, 2019; Kuo, 2015; Liu, 2011), Amis and its preservation have not yet benefited from (to the best of our knowledge) data-based methods used in machine learning and natural language processing.

Low-resource machine translation has recently attracted more attention in the field of natural language processing for languages such as Amis that have a relatively low amount of data due to a small population of speakers. Because neural machine translation (NMT) systems typically do not perform well for low-resource languages, which lack parallel data (Koehn and Knowles, 2017), approaches such as collaborating with language communities to increase parallel data, transfer learning from other machine translation systems, and using multilingual models, among others are being explored (Haddow et al., 2022). However, despite all these new approaches to low-resource machine translation, it is clear that parallel data is still essential for training state-of-the-art machine translation systems (Haddow et al., 2022), as high-resourced language pairs still require large amounts of data to achieve state-of-the-art translation quality (Akhbardeh et al., 2021).

Due to the lack of Amis resources available for use in machine translation, we developed an Amis-Mandarin parallel corpus and dictionary. Our contributions can be summarized as follows:

- We present an Amis-Mandarin dataset, which consists of an Amis-Mandarin parallel corpus containing 5,751 sentences and a dictionary containing 7,800 unique words and phrases in Amis with definitions in Mandarin.
- We trained neural machine translation models on the Amis-Mandarin dataset and produced baselines for future studies.

2 Amis

Amis (ISO 639-3 language code *ami*) is an East Formosan language (Blust, 1999; Ross et al., 2009) spoken on the east coast of Taiwan between Hualien and Taitung (Liu, 2011; Kuo, 2015) by the Amis, one of Taiwan's several indigenous ethnic groups. Formosan languages are spoken by the indigenous peoples of Taiwan (Liu, 2011) and are part of the Austronesian language family. They are believed to be the most diverse of the Austronesian languages (Li, 2008), and because high diversity in a group of genetically-languages found in a geographical area implies earlier settlement in that area (Sapir, 1916), Taiwan is considered to be the homeland of Austronesian languages (Li, 2008; Blust, 1999). Figure 1 shows the distribution of these languages in Taiwan with the region where Amis is spoken being shaded in gray.

Though the vast majority of Taiwan's population is Han Chinese, Taiwan is home to several groups of indigenous peoples who are Austronesian (Trejaut et al., 2014). According to official

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 79–84
November 20, 2022.

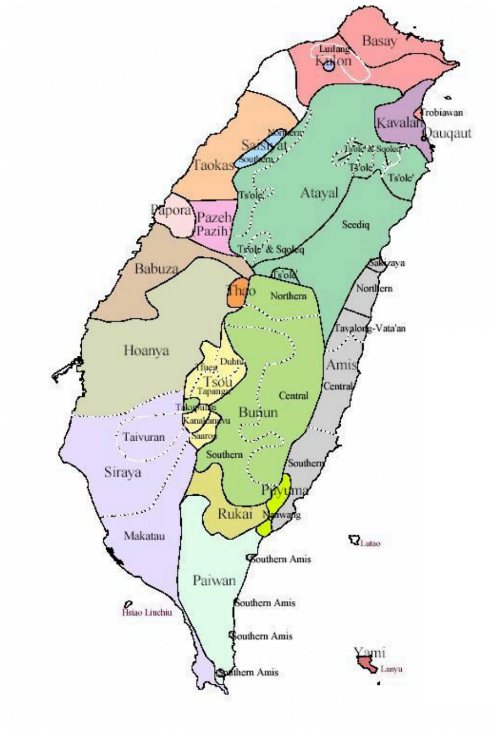

Figure 1: Distribution of Taiwan's Indigenous Languages (Li, 2004, as cited in Liu, 2011).

government statistics[1], Taiwan's indigenous population is 582,008, which is approximately 2.4% of Taiwan's total population. The Amis people have a population of 217,216, making up approximately 37.3% of Taiwan's indigenous population. Mandarin Chinese is the language of education (Scott and Tiun) and is spoken along with other Chinese languages (e.g. Hokkien) by the majority of the population, whether indigenous or not. Despite the Amis population being over 200,000, the Amis language has just roughly 30,000 speakers (Kuo, 2015).

These roughly 30,000 speakers, however, do not all speak the same dialect of Amis. According to Tsuchida (1982, 1988, as cited in Kuo, 2015), Amis has five major dialects: (i) Sakizaya (撒奇萊雅群), (ii) Northern/Nanshi Amis (北部/南勢阿美群), (iii) Tavalong-Vata'an (太巴塱-馬太鞍群), (iv) Central/Haian Amis (中部/海岸阿美群), and (v) Southern/Peinan and Hengchun Amis (南部/卑南恆春阿美群). The dataset described in Section 3 in this paper uses data from Central Amis. The existence of several dialects of Amis means that any one dialect has a relatively low number of speakers.

Amis is classified as a vulnerable language by UNESCO (Moseley, 2010), meaning that most, but not all, children and families of the Amis, an indigenous Austronesian ethnic group native to Taiwan, speak Amis as a first language but that their use of Amis may be limited to specific social settings (such as the home, where it is used amongst family) (Moseley, 2010). However, Amis has also been described as an endangered language by several linguists (Edmondson et al., 2005; Kuo, 2015; Liu, 2011). Liu (2011), who researched the Amis language in Taiwan since 1995, performing extensive fieldwork and data gathering with native Amis speakers, notes that only those over 50 are proficient in Amis. Whether Amis is "vulnerable" or "endangered," it is clear that the language is at risk and would benefit from more attention from linguists and natural language processing technologies to help preserve the language digitally and enable its use in modern technologies.

3 Amis-Mandarin Dataset

We compiled Amis-Mandarin parallel data from an Amis-Mandarin online dictionary (原住民族語言線上辭典)[2] published by the Indigenous Languages Research and Development Foundation (原住民族語言研究發展基金會). This dictionary consists of words, phrases, and example sentences of these words and phrases in Central Amis with their Mandarin translations. It was compiled in 2012 by National Taiwan Normal University.

Though these data have been made searchable in the format of an online dictionary, they are not designed for computational use. The website allows one to download the dictionary in parts or in whole as a PDF or ODT file. However, due to some inconsistencies in how words, translations, and example sentences are laid out in these files, neither is easy to use for computational tasks. Thus, we downloaded PDFs of the dictionary made available by this online dictionary, converted them to HTML using PDFMiner[3], and extracted data using Beautiful Soup[4] indepen-

[1] July 2022 statistics from Taiwan's Council of Indigenous Peoples (原住民族委員會) https://www.cip.gov.tw/zh-tw/news/data-list/940F9579765AC6A0/C89C009B11A070EC725C4C571E9FFD7B-info.html

[2] https://e-dictionary.ilrdf.org.tw/ami/search.htm

[3] https://github.com/pdfminer/pdfminer.six

[4] https://www.crummy.com/software/

Table 1: Summary of the Amis-Mandarin Parallel Corpus

	Total	Train	Dev	Test
Number of sentences	5,751	4,600	576	575
Number of Amis words	38,946	31,136	3,947	3,863
Number of Chinese characters	69,864	55,672	7,289	6,903

dently from the authors of this dictionary. Each Amis word/phrase and its Mandarin dictionary entry were extracted, and when available, example sentences in Amis along with their Mandarin translations were also taken. Dictionary entries and their Mandarin definitions were put into one pickle file, while example sentences and their Mandarin translations were put into another pickle file. These files can be opened using pandas[5]. The dictionary is also available as a tab-delimited text file, and the parallel sentence data are also available as text files split into train, dev, and test sets. The parallel data were shuffled before being split into the train, dev, and test sets, which were taken from 80%, 10%, and 10% respectively from the shuffled data. The dataset we compiled can be found at https://github.com/francisdzheng/amis-mandarin.

Our dictionary dataset contains 7,800 unique entries in Amis along with their Mandarin equivalents. Amis dictionary entries that had more than one definition in Mandarin were added to our dictionary dataset as separate entries for each additional definition. Thus, there are a total of 7,926 pairs of Amis and Mandarin words/phrases in our dictionary dataset. Our parallel corpus dataset contains 5,751 Amis sentences and their Mandarin translations. This parallel corpus dataset is summarized in Table 1. Due to the concept of a word being different in Amis and Mandarin, Table 1 describes the Mandarin data in terms of characters and the Amis data in terms of words, which are typically separated by spaces unlike in Mandarin, which does not use spaces in writing.

4 Amis-Mandarin Machine Translation

Using our Amis-Mandarin dataset, we trained models for machine translation between Amis and Mandarin in both directions.

4.1 Methods

4.1.1 Preprocessing

Data were tokenized using a unigram (Kudo, 2018) implementation of SentencePiece (Kudo and Richardson, 2018). A vocabulary size of 4,000 and a character coverage rate of 0.9995 were used. Using our SentencePiece (Kudo and Richardson, 2018) model and vocabulary, we used FAIRSEQ[6] (Ott et al., 2019) to build vocabularies and binarize our training data in preparation for training our model.

4.1.2 Training

We trained a Transformer (Vaswani et al., 2017) model using an mBART (Liu et al., 2020) implementation of FAIRSEQ (Ott et al., 2019) for translation between Amis and Mandarin in both directions. Our Transformer (Vaswani et al., 2017) model used six encoder and decoder layers with eight attention heads each, a hidden dimension of 512, and a feed-forward size of 2048, and a learning rate of 0.0003. Our model was optimized using Adam (Kingma and Ba, 2015) with hyperparameters $\beta = (0.9, 0.98)$ and $\epsilon = 10^{-6}$. A dropout rate of 0.1 and a weight decay of 0.01 were used for regularization.

We conducted two experiments, one in which training involved only the training set from our parallel Amis-Mandarin corpus (consisting of sentences) and one which included the dictionary dataset as part of the training data. The dictionary data were treated as additional parallel data (though they're not full sentences) and simply added on to the parallel sentence training data. This was done to see the effect of exposing models to the dictionary dataset and to establish two baselines for translation as a dictionary may not always be available when training models.

4.1.3 Evaluation

Translations outputted by our model were evaluated with detokenized BLEU (Papineni et al.,

BeautifulSoup/

[5] https://pandas.pydata.org

[6] https://github.com/facebookresearch/fairseq

Table 2: Results

	Without Dictionary		With Dictionary	
	BLEU	**CHRF**	**BLEU**	**CHRF**
Amis to Mandarin	5.33	0.1596	**7.07**	**0.2198**
Mandarin to Amis	15.36	0.4018	**18.94**	**0.4618**

2002; Post, 2018) using the SacreBLEU library[7] (Post, 2018) on the test data from our parallel corpus. We also used CHRF (Popović, 2015) to measure performance at the character level.

4.2 Results

Our results are presented in Table 2. Models trained using only the parallel corpus dataset performed worse than the models trained using both our parallel corpus dataset and dictionary dataset. This is expected as these models were able to learn direct translations of individual words and phrases that are used in the parallel sentence data in addition to translations of whole sentences. Though dictionaries are not as useful as parallel data in that dictionaries do not reveal much about how a word or phrase should be used in a sentence, using dictionary data in the training process proved to significantly improve translation quality.

The improvement in translation quality after adding the dictionary dataset can be seen in both the Amis → Mandarin and Mandarin → Amis directions and is reflected in both the BLEU and CHRF scores. Notably, the improvement in translation quality as measured by BLEU for the Mandarin → Amis direction was greater than that for the Amis → Mandarin direction. One possible explanation for this is the fact that some Amis words in the dictionary dataset are paired with multiple Mandarin equivalents or longer Mandarin explanations, exposing the model to relatively more Mandarin words for a single given word or phrase in Amis. Thus, the model may map multiple words in Mandarin to single words or phrases in Amis, which are almost sure to appear in the parallel sentence data (as mentioned in Section 3, the parallel sentence data come from example sentences for the Amis words in the dictionary). On the other hand, the Mandarin definitions for each Amis entry in the dictionary dataset do not necessarily appear in the parallel sentence data. More research is needed to see whether the model trained on both the parallel corpus dataset and dictionary dataset still performs better in the Mandarin → Amis direction on other parallel data.

[7] https://github.com/mjpost/sacrebleu

5 Conclusion

We presented an Amis-Mandarin parallel corpus and dictionary, which is the first, to the best of our knowledge, Amis-Mandarin dataset documented in English for the natural language processing community. Though the online dictionary from which we obtained the data is available to anyone, the dictionary interface is only available in Mandarin, and the data is not in a form that NLP researchers can easily use. Other Amis-Mandarin data we found on the web were also not in an easily usable format and not friendly for English-speaking researchers. The dataset we compiled consists of 5,751 parallel sentences and 7,800 Amis words and phrases paired with their definitions in Mandarin. Using this dataset, we experimented with Amis-Mandarin machine translation and established baseline BLEU and CHRF scores. Using both the parallel corpus and dictionary during training produced models that performed the best on our test data from our parallel corpus.

Aside from sentence translation, we also envision that the dataset we compiled can be used for exploring how dictionary entries can be predicted using parallel data (instead of using dictionary entries to aid in the translation of sentences). Defining words is also an important part of language documentation, and it would be interesting to see how machines can draw meaning for individual words or short phrases given parallel sentence data.

In the future, we would like to take a closer look at how tokenization can be optimized for the two languages and try using other existing tokenizers that have been trained more specifically for Mandarin. We also want to try to incorporate more external knowledge and perhaps acknowledge that parallel data may never be enough. Our dataset is small, and we hope to explore how knowledge

from grammars and other literature written on Amis can be incorporated into our model or into the creation of synthetic parallel data. As Amis is an Austronesian language like Indonesian, which is widely spoken and has much more literature available, it is possible that knowledge from Indonesian can be helpful in NLP tasks involving Amis. We hope that our dataset can spark more interest from the machine learning community in not only Amis, but other Formosan languages as well.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Robert Blust. 1999. Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. In E. Zeitoun and Paul Jen-Kuei Li, editors, *Selected Papers from Eighth International Conference on Austronesian Linguistics*, pages 31–94. Academica Sinica, Taipei.

Jerold Edmondson, John Esling, Jimmy Harris, and Tung-Chiou Huang. 2005. A laryngoscopic study of glottal and epiglottal/pharyngeal stop and continuant articulations in Amis– an Austronesian language of Taiwan. *Language and Linguistics*, 3:381–396.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. pages 1–60.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jonathan Cheng-Chuen Kuo. 2015. *Argument alternation and argument structure in symmetrical voice languages: A case study of transfer verbs in Amis, Puyuma, and Seediq*. Ph.D. thesis, University of Hawai'i at Manoa.

Paul Jen-kuei Li. 2008. The great diversity of Formosan languages. *Language and Linguistics*, 9(3):523–546.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Tsai-hsiu Liu. 2011. *Complementation in three Formosan languages: Amis, Mayrinax Atayal and Tsou*. Ph.D. thesis, University of Hawai'i at Mnoa, Honolulu, HI.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*,

pages 392–395, Lisbon, Portugal. 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.

Malcolm Ross et al. 2009. Proto Austronesian verbal morphology: A reappraisal. In *Austronesian historical linguistics and culture history: A festschrift for Robert Blust*. Asia-Pacific Linguistics, College of Asia and the Pacific, The Australian ⋯.

Edward Sapir. 1916. *Time perspective in aboriginal American culture: A study in method*. 13. Government Printing Bureau.

Mandy Scott and Hak-khiam Tiun. Mandarin-only to Mandarin-plus: Taiwan. 6(1):53–72.

Jean A. Trejaut, Estella S. Poloni, Ju-Chen Yen, Ying-Hui Lai, Jun Hun Loo, Chien liang Lee, Chunfen He, and Marie Lin. 2014. Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. *BMC Genetics*, 15:77 – 77.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Li-ying Wu and Ken Lau. 2019. Language education policy in Taiwan. In *The Routledge international handbook of language education policy in Asia*, pages 151–161. Routledge.