Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration

Ryo Igarashi
n33t5hin@gmail.com

So Miyagawa
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Ibaraki, Japan
miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

Neural Machine Translation (NMT) has revolutionized language translation, yet significant challenges persist for low-resource languages, particularly those with high dialectal variation and limited standardization. This comprehensive study focuses on the Ainu language, a critically endangered indigenous language of northern Japan, which epitomizes these challenges.

We address the limitations of previous research through two primary strategies: (1) extensive corpus expansion encompassing diverse domains and dialects, and (2) development of innovative methods to incorporate dialect and domain information directly into the translation process. Our approach yielded substantial improvements in translation quality, with BLEU scores 39.06 for Japanese → Ainu and 31.83 for Ainu → Japanese.

Through rigorous experimentation and analysis, we demonstrate the crucial importance of integrating linguistic variation information in NMT systems for languages characterized by high diversity and limited resources. Our findings have broad implications for improving machine translation for other low-resource languages, potentially advancing preservation and revitalization efforts for endangered languages worldwide.

1 Introduction

Ainu is the indigenous language of the Ainu people, who are native to northern Japan, Sakhalin, and the Kuril Islands.

Due to the Japanese government's assimilation policy during the 20th century, the number of people speaking Ainu as their first language drastically declined. Today, UNESCO classifies Ainu as a critically endangered language, and estimates suggest that fewer than ten native speakers remain, all of whom are elderly (Moseley, 2010).

However, there has been a growing focus on revitalizing the Ainu language in recent years. This development follows the Japanese government's official recognition of the Ainu as an indigenous people, which has led to national funding for Ainu language courses and educational materials (Sato, 2012).

Many Ainu learners speak Japanese as their first language today; thus, practical machine translation is integral to the revitalization of Ainu. However, a previous study by Miyagawa (2023) faced significant challenges, including difficulties in distinguishing between different dialects and challenges in translating everyday conversation.

To address these problems, we carried out the following approaches.

Firstly, we enhanced the corpus. Previous studies' corpora were predominantly biased toward folklore from limited regions. We gathered and digitized resources from various dialects and domains to ensure greater diversity.

We also introduced a novel approach to Ainu-Japanese translation that can distinguish dialects and domains, reducing wording confusion between different regions or contexts.

In this paper, we elaborate on the details of the methodology, present our results, and discuss the implications of our findings that can potentially contribute to the revitalization of Ainu, which may also apply to other low-resource languages.

2 Background and Related Work

In this section, we will review the background of the Ainu language and discuss previous studies.

2.1 The Ainu Language

Ainu is a language isolate with no demonstrable genetic relationship to any other languages, including neighboring languages such as Japanese.

Furthermore, Ainu is a polysynthetic language, where complex words with extensive meanings can be created by combining multiple affixes (Tamura, 2020).

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 413–422
November 16, 2024.

Additionally, Ainu does not have a native writing system, and currently, it is written using the Latin alphabet or Katakana. In particular, the orthography used in the textbook *AKOR ITAK*, published by the former Hokkaido Utari Association, has been broadly accepted by learners and adopted in other publications (Nakagawa, 2006).

2.2 Challenges in Ainu Language Processing

Neural language processing in Ainu faces several significant challenges.

Firstly, the Ainu language is not standardized, leading to regional variations in expressions. These differences are widespread and affect vocabulary, grammar, and pronunciation (Hattori and Chiri, 1960).

Secondly, expressions in Ainu vary significantly depending on context. In Ainu, vocabulary and wording change based on whether the language is used in storytelling, such as folklore narration, or in everyday conversation. One notable difference is the use of personal affixes reflecting logophoricity (Bugaeva, 2008). For example, in conversation, actions of a speaker are marked by a first-person prefix *ku*=, but in folklore, it often changes to a fourth-person affix *a*= or =*an*. Conversely, using *a*= or =*an* in conversation indicates a quotation or an inclusive "we" (Nakagawa, 2011).

Lastly, the availability of corpora for Ainu language processing is extremely limited. Although institutions including the National Institute for Japanese Language and Linguistics (NINJAL) and the National Ainu Museum have made efforts to collect and digitize some corpora, the overall quantity remains insufficient. Moreover, most of these corpora focus on folklore from specific regions, which does not adequately capture the full diversity of the Ainu language.

Therefore, it is essential to expand the corpus to include resources from various domains and regions and to enhance the machine translation model's ability to handle ambiguities. These steps are crucial for improving the performance of natural language processing in Ainu.

2.3 Previous Work in Ainu-Japanese Machine Translation

Ptaszynski et al. (2013) proposed an initial implementation of rule-based Ainu-Japanese machine translation. This system internally uses a part-of-speech tagger based on a Hidden Markov Model, replacing each Ainu word with its Japanese equivalent. This approach works well for Japanese, where word order is closely aligned with Ainu.

Furthermore, Miyagawa (2023) experimented with Transformer-based neural machine translation in Ainu-Japanese, achieving a BLEU score of 32.90 for Japanese-to-Ainu and 10.45 for Ainu-to-Japanese, the highest reported to date for this task.

However, it is important to note that both studies trained their models using folklore from a specific region, which limits their applicability to conversations or other dialects.

3 Methodology

This section details our methodology for selecting materials in the corpus, preparing them in a format suitable for machine learning, and conducting the training process.

3.1 Corpus

As mentioned above, most digitized corpora currently available are lacking in diversity, primarily containing folklore from limited regions. To address this, we collected additional resources to enhance the diversity of our corpus.

In addition to already digitized resources, we established the following criteria to prioritize resources for digitization:

- **Writing System:** We exclusively selected resources written in the Latin alphabet according to the *AKOR ITAK* orthography. Although Katakana is also widely used, we chose not to include it due to its lower accuracy in Optical Character Recognition (OCR), which is essential for converting printed text into a digital format. Additionally, while it is possible to convert Latin scripts into Katakana, the reverse is not feasible, as Katakana does not distinguish between phonemes such as i and y or u and w. Moreover, Katakana does not clearly indicate the boundaries between personal affixes and other word components.

- **Wider Variety:** We prioritized resources that contain modern texts, conversations, and less documented dialects to ensure a comprehensive and representative dataset that reflects the full range of linguistic diversity in Ainu.

As a result of an extensive review of available books, websites, textbooks, dictionaries, and peri-

odicals, we successfully collected a diverse set of resources, which are detailed in Appendix A.

3.2 Preprocessing

Most of the resources we selected are not machine-readable. Even when they were available as PDF files, we needed to establish the correspondence between Ainu sentences and their counterpart Japanese translations. We addressed this issue through the following steps.

Firstly, we scanned the printed materials and used the Cloud Vision API[1] to perform OCR. This allowed us to obtain machine-readable text data along with metadata, including dimensions and coordinates, similar to the metadata structure typically found in PDFs. Although the Cloud Vision API does not officially support the Ainu language, it has basic recognition capabilities for the Latin alphabet. Therefore, we adapted it with additional validation, as described in the following step.

Next, we developed a Node.js script to establish the correspondence between Ainu text and Japanese translations, aligning the parallel corpus. This script utilized the dimensions and coordinates obtained in the previous step. We adjusted the threshold configurations for alignment based on the layouts of each resource.

Finally, we validated the obtained parallel corpus. As expected, the Cloud Vision API, which does not support the Ainu language, produced various recognition errors. These errors included, but were not limited to, incorrect recognition of personal affixes (e.g., *k=arpa* recognized as *karpa*) and issues with single-character words being joined with their sibling words (e.g., *ye p* recognized as *yep*). To address these errors, we validated the recognized text using the customizable open-source spell checker Code Spell Checker[2], with a word list extracted from the dictionaries available on the Ainu Language Archive[3]. We then conducted a thorough manual review to correct any remaining recognition errors.

3.3 Format

In this section, we outline the corpus format developed for this study. While alternative methods could offer advantages from the perspective of indexing, we prioritized simplicity, given that our primary objective was developing a neural machine translation model.

3.3.1 Domains

One of the significant challenges we faced was categorizing the domains of the collected resources.

We initially considered Nowakowski et al.'s (2018) approach, which classified existing corpora into 15 distinct genre types. However, some of our newly collected resources lacked the necessary detail for categorization within the context of Ainu folklore. Accurate classification would have required a deep understanding of Ainu language and culture.

Therefore, we opted against a nuanced categorization of all resources and instead adopted a more straightforward and objective approach: utilizing the personal affix predominantly used in a resource as a proxy for domain categorization. This approach allowed us to automatically classify the resources into two mutually exclusive classes: first-person and fourth-person.

3.3.2 Dialects

Designing maintainable classes for dialects also presented significant challenges due to the lack of consensus on how to classify the various Ainu dialects.

Hattori and Chiri (1960), a pioneering study of Ainu dialects, lists 19 dialects, while a subsequent survey by Asai (1974) expanded this to 21, adding three additional dialects.

As a result, the dialect names used in different resources are inconsistent. For instance, the Ainu Times does not distinguish between the Saru and Chitose regions, instead referring to them collectively as "Saru-Chitose".

Given these inconsistencies, we ultimately decided to retain the dialect names as they were listed in each resource. Consequently, our corpus includes instances labeled as "Saru", "Saru-Chitose", and "Western Hokkaido". While this approach may result in overlapping categories, we did not consider this issue critical, as we hypothesized that modern language models would be capable of discerning the similarities and differences between these dialects.

Finally, we consolidated all the data into the format shown in Table 1 and compiled it using the Hugging Face Datasets library. This process resulted in the creation of a novel corpus (Table 2)

[1] https://cloud.google.com/vision
[2] https://cspell.org/
[3] https://ainugo.nam.go.jp/

comprising approximately 1.2 million words and 4.2 million characters.

3.4 Normalization

We applied minimal normalization, which involved removing diacritics (e.g., húre → hure), linking symbols (e.g., or_un → or un), and footnote markers inserted by editors.

3.5 Training

In this section, we provide a detailed explanation of the training setup.

3.5.1 Model

Lee et al.'s (2022) prior research demonstrated the effectiveness of multilingual sequence-to-sequence language models, such as mT5 (Xue et al., 2020) and mBART (Chipman et al., 2022), for handling machine translation tasks in low-resource languages. For the present study, we selected mT5 due to the extensive research supporting its performance.

Firstly, to reduce costs, we trained the `mt5-small` model, which features a relatively small number of parameters, using various task prefixes, as described in the following section. This preliminary step allowed us to confirm that including metadata, such as dialects and domains, contributes to improved performance.

Next, using the most effective task prefix identified in the previous step, we trained and compared the `mt5-base` and `t5-base` models. This comparison aimed to assess both the effectiveness of the multilingual model and the impact of the number of parameters on performance.

Finally, we conducted additional training with the `mt5-base` model using only specific domain-dialect pairs to evaluate the effectiveness of mixing different types of resources within the corpus. In this step, we focused on folklore and conversations from the Saru, Chitose, Shizunai, and Horobetsu regions, where relatively abundant data are available.

For practical reasons, we trained both the Ainu-to-Japanese and Japanese-to-Ainu translation tasks within the same model.

3.5.2 Task Prefix

T5 and mT5 support multi-task learning by embedding tokens known as task prefixes in an input sequence. This approach enhances overall performance by allowing the model to differentiate between various task types (Raffel et al., 2020).

We hypothesized that applying this method to Ainu could improve machine translation performance by effectively disambiguating subtle linguistic differences across dialects and domains.

To test this hypothesis, we conducted experiments using the four variations of language names shown in Table 3.

Here, `dialect` refers to the specific dialect of the resource, and `domain` refers to the predominant personal affix used in the text. For example, if translating an Ainu folklore text written in the Saru dialect to Japanese, the P_{both} prefix would be `translate Ainu (Saru, fourth) to Japanese`. For translations in the opposite direction, the task prefix would be adjusted accordingly to `translate Japanese to Ainu (Saru, fourth)`.

3.5.3 Settings

The training process was conducted using the Hugging Face Transformers library. To orchestrate the infrastructure and ensure consistent metric measurement within the same environment, we developed a training pipeline using Google Cloud's Vertex Pipelines.

Hyperparameter tuning was performed using Vertex AI's Hyperparameter Tuning Job. We utilized 10% of the dataset to perform a grid search for optimal hyperparameters, as outlined in Table 4.

3.5.4 Evaluation

We employed stratified sampling (Japkowicz and Stephen, 2002) by dialect-domain pairs to split the dataset proportionally. Specifically, 10% of the corpus was allocated as the evaluation set, while the remaining 90% was used for training.

Evaluation metrics were calculated using the BLEU score (Papineni et al., 2002), with SacreBLEU, a commonly used library in prior research, employed for its computation. For Japanese-to-Ainu translations, we used the default `13a` tokenizer, and for Ainu-to-Japanese translations, we used the `ja-mecab` tokenizer.

4 Results

In this section, we elaborate on the results of the experiments.

4.1 Task Prefixes

Here, we examine how performance varies depending on different task prefixes. Table 5 shows the

Name	Type	Description
`book`	`str`	Book title
`title`	`str`	Title of the text
`domain`	enum of `"first"` and `"fourth"`	Type of personal affix of the speaker
`author`	`Option[str]`	Author of the text (if known)
`dialect`	`Option[str]`	Dialect of the text (if known)
`text`	`str`	Sentence in Ainu
`translation`	`str`	Translation in Japanese

Table 1: Corpus format

Title	Type	Words	Characters
The Ainu Language Archive	Web	600,770	2,107,984
The Ainu Times	Book	148,843	519,040
Collection of Ainu Oral Literature	PDF	135,649	492,484
ILCAA Online Text of Ainu Collected by Suzuko Tamura	Web	95,379	299,630
A Glossed Audio Corpus of Ainu Folklore	Web	76,550	243,696
Dictionary of the Mukawa dialect of Ainu	Web	66,386	247,637
Ainu textbooks by The Foundation for Ainu Culture	PDF	25,067	84,905
Bulletin of the Hokkaido Ainu Culture Research Center	PDF	14,724	48,092
A Topical Dictionary of Conversational Ainu	Web	13,831	49,776
Ainu Shin'yōshū	Book	10,364	38,153
New Express Plus Ainu-go	Book	4,418	14,812
Learning Ainu Language by Listening to Kamuy Yukar	Book	3,028	11,177
AKOR ITAK	Book	2,005	5,903
Total		**1,197,014**	**4,163,289**

Table 2: The Ainu language resources. We treated personal affixes as separate words and excluded line breaks and whitespace from the character count.

Label	Language Name in Task Prefix
P_{none}	`Ainu`
P_{dialect}	`Ainu (dialect)`
P_{domain}	`Ainu (domain)`
P_{both}	`Ainu (dialect, domain)`

Table 3: The list of strings we used as a language name for the Ainu language.

BLEU scores for each task prefix.

With P_{none}, we observed a BLEU score of 29.89 for Ainu-to-Japanese and 32.24 for Japanese-to-Ainu, which does not significantly differ from the results reported by Miyagawa (2023).

However, with P_{domain}, we observed a subtle performance improvement, with scores of 29.93 for Ainu-to-Japanese and 32.70 for Japanese-to-Ainu.

With P_{dialect}, performance improved significantly for Japanese-to-Ainu, with a BLEU score of 35.94. We also observed a slight improvement for Ainu-to-Japanese, with a score of 30.40.

Finally, with P_{both}, we achieved the highest performance, with BLEU scores of 30.70 for Ainu-to-Japanese and 36.25 for Japanese-to-Ainu.

4.2 Models

We also experimented with different sequence-to-sequence models to determine which one performs best. Given that P_{both} was proven to be the most effective task prefix, all models were trained using this prefix. Table 6 shows the BLEU scores for each model.

With `mt5-base`, performance improved for both translation directions, achieving BLEU scores of 31.83 for Ainu-to-Japanese and 39.06 for Japanese-to-Ainu, making it the best-performing model among all the models tested.

In contrast, the `t5-base` model failed to produce practical translation results, with BLEU scores of 0.00 for Ainu-to-Japanese and 0.01 for Japanese-to-Ainu. As these scores indicate, the model generated nothing but nonsensical text.

Parameter	Value
Framework	Hugging Face Transformers (v4.40.1)
Infrastructure	Google Cloud Vertex AI
Hardware	`a2-highgpu-1g` instance with NVIDIA A100 GPU
Scheduler	Linear scheduler with 6% warm-up steps
Learning rate	Maximum of 5.0×10^{-5}
Optimizer	AdamW with weight decay of 1.0×10^{-3}
Context size	128 tokens
Batch size	16 with gradient accumulation every 2 steps
Training duration	Maximum of 20 epochs with early stopping (patience=3)

Table 4: Training settings

Task Prefix	ain→ja	ja→ain
P_{none}	29.89	32.24
P_{domain}	29.93	32.70
P_{dialect}	30.40	35.94
P_{both}	30.70	36.25

Table 5: Ainu-Japanese translation performance for each task prefix

Model	ain→ja	ja→ain
`t5-base`	0.00	0.01
`mt5-small`	30.70	36.25
`mt5-base`	31.83	39.06

Table 6: Ainu-Japanese translation performance for each model

4.3 Performance for Each Domain and Dialect

We also evaluated the performance of the `mt5-base` model across different domains and dialects. Table 7 compares the performance of the mixed corpus model with that of a model trained exclusively on specific domain-dialect pairs.

Across all classes, the model trained on the mixed corpus consistently outperformed the model trained on individual classes. Notably, we observed significant improvements in classes with smaller datasets, such as conversations in the Chitose or Horobetsu dialects.

5 Discussion

This section discusses how multilingual pre-trained models, domain, and dialectal variations impact Ainu machine translation.

5.1 Effectiveness of Domain and Dialect

This study confirmed that incorporating dialectal information significantly improves translation performance. This improvement is likely due to the linguistic variations that exist across different regions. Notably, greater performance gains were observed in the Japanese-to-Ainu translation. This difference may be due to the ambiguity in determining which dialect to use as the target when translating from Japanese. By specifying the target dialect, this ambiguity is resolved, allowing the model to produce more accurate translations and resulting in improved performance.

Here is an example of translating "I go to the mountain to pick mushrooms" (*Watashi wa kinoko o tori ni yama e ikimasu.*) to different dialects. Note that the model correctly used the appropriate wording for each dialect:

Saru: karus ku=kar kusu **ekimne** k=**arpa**.
Tokachi: karus ku=kar kusu **ekimun** k**u=oman**.

Our study also found that the inclusion of domain metadata led to performance improvements, although these gains were less pronounced compared to those achieved with dialectal information. One possible reason for this difference could be the complexity of the vocabulary and unique expressions found in folklore, which may have posed challenges for the model.

Here is an example of translating "I want to eat with my friend" (*Tomodachi to issho ni shokuji shitai.*) in different domains. Note that the model correctly adjusted the personal affixes according to the domain:

Folklore: **a=**utari turano ipe**=an** rusuy.
Conversation: **k=**utari turano **ku=**ipe rusuy.

We believe this approach could also be applicable to other languages that lack standardization, especially endangered languages with context and dialect variations. For instance, the Ryukyuan languages, characterized by an extensive politeness

Dialect	Domain	Words	Exclusive Corpus		Mixed Corpus	
			ain→ja	ja→ain	ain→ja	ja→ain
Saru	Conversation	25,506	23.05	33.03	35.47	42.94
	Folklore	527,728	24.48	32.14	28.86	33.31
Shizunai	Conversation	24,403	22.37	37.57	37.48	50.74
	Folklore	233,134	36.35	44.16	38.81	47.31
Chitose	Conversation	6,487	13.51	13.92	70.97	70.96
	Folklore	15,664	12.02	27.51	36.88	42.36
Horobetsu	Conversation	4,382	0.22	0.36	89.19	80.89
	Folklore	10,364	1.23	2.19	30.56	38.21

Table 7: Ainu-Japanese translation performance metrics for each domain and dialect

system and numerous dialects influenced by the archipelagic geography, present a similar challenge. Embedding politeness levels and regional information in the task prefix could improve MT performance for these languages by providing more accurate and contextually appropriate translations. The model might better manage linguistic nuances and variability by explicitly incorporating such metadata, enhancing translation accuracy.

5.2 Advantages of Multilingual Pre-Trained Models

Building on previous studies, our research confirms the applicability of multilingual pre-trained language models for Ainu-Japanese translation. This finding supports the use of these models for low-resource languages and demonstrates the potential of transfer learning, even for a language isolate such as Ainu.

However, these models should not be considered a universal solution, as they have several drawbacks. One significant disadvantage is the large number of parameters required by mT5, which has a considerably larger vocabulary size to handle tokens from multiple languages. While this enables them to provide strong baseline performance across various language tasks, it also results in excessive parameters for specific tasks, such as translation between particular languages. This leads to inefficiencies and requires substantial computational resources for both training and inference.

Additionally, the Sentencepiece tokenizer (Kudo and Richardson, 2018) used in mT5 was not specifically trained on Ainu texts, leading to suboptimal tokenization. For example, basic sentences such as *irankarapte. tanto sirpirka wa.* are tokenized into eleven separate tokens, with even fundamental words including *pirka* being split unnecessarily:

```
['_ir', 'ankara', 'pte', '.', '_',
'tanto', '_sir', 'pirk', 'a', '_wa', '.']
```

5.3 Impact of Mixing Multiple Dialect-Domain Classes on Model Performance

Our study found that training a model by mixing multiple domain-dialect classes and using task prefixes to distinguish them results in higher performance compared to training on a single domain-dialect class. This effect is particularly pronounced in classes with limited resources. This finding suggests that there is shared grammar or vocabulary among different classes, which a language model can leverage to enhance performance when needed.

This finding could also be valuable for other endangered languages where collecting more materials for a specific dialect is impractical. Our results demonstrate that MT performance can still be enhanced by incorporating resources from other dialects and distinguishing them using task prefixes. This approach allows the model to make use of shared linguistic features across dialects, effectively broadening the usable data pool and compensating for individual dialect resource limitations.

6 Conclusion

This study demonstrates the effectiveness of incorporating dialect and domain information in NMT systems for low-resource languages with high linguistic variation. By expanding the corpus and leveraging task prefixes to provide contextual information, we achieved significant improvements in Ainu-Japanese translation performance.

Our research contributes to the broader field of low-resource language NMT by:

1. Highlighting the importance of diverse, well-annotated corpora.

2. Demonstrating the potential of integrating linguistic metadata into the translation process.

3. Providing insights into the challenges and strategies for addressing languages that lack standardization and exhibit high variation.

As we continue to refine and expand these methods, we move closer to achieving effective machine translation for all languages, irrespective of their resource status. This work not only advances NMT research but also contributes to broader language preservation and revitalization efforts, providing new tools and methodologies for engaging with endangered languages.

Limitations

While our research achieved the highest scores to date in Ainu-Japanese machine translation, several limitations affect the generalizability and performance of our model.

Limited Corpus

The most significant limitation is the insufficient amount of the Ainu language data. Despite our efforts to digitize and format the most comprehensive Ainu corpus currently available, the dataset remains too limited for extensive machine translation training.

This scarcity of data is particularly noticeable in the lack of folklore from regions other than Saru and Shizunai, as well as a shortage of conversational resources across all dialects.

Additionally, some existing resources could not be utilized due to inconsistencies in writing systems. Developing a model that can convert between different writing systems may help address this issue. While data augmentation methods, such as back-translation, could be used to expand the corpus, their effectiveness is likely limited, as most existing Ainu resources come with Japanese translation.

Furthermore, we decided not to make our collected corpus publicly available due to copyright restrictions. This limitation poses challenges for performance comparisons in future research. A model trained exclusively on specific domain-dialect pairs could potentially achieve a higher BLEU score than our model, but this would not necessarily indicate superior performance across the broader spectrum of the Ainu language. Establishing more consensus on the digitization and use of copyright-protected works, particularly for endangered languages, could help address this issue and facilitate broader research efforts.

Fine-Grained Dialects and Domains

In this study, we employed a simplified approach to classify domains and dialects. Consequently, our model cannot perform translations that target more specific regions or dialects. For example, there are different types of folklore, such as *yukar* and *uwepeker*, both of which are narrated using the same fourth-person affix. Our current approach does not differentiate between these types.

Future research would benefit from a more finely annotated corpus, particularly with respect to capturing subtle differences in dialects or domains.

References

Toru Asai. 1974. *Classification of dialects: Cluster analysis of Ainu dialects*. Institute for the Study of North Eurasian Cultures, Hokkaido University.

Anna Bugaeva. 2008. Reported discourse and logophoricity in southern hokkaido dialects of ainu. *GENGO KENKYU (Journal of the Linguistic Society of Japan)*, 133:31–75.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. 2022. mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Shirō Hattori and Mashiho Chiri. 1960. Ainu-go shuhōgen no kiso goi tōkeigakuteki kenkyū [a lexicostatistical study of basic vocabulary in Ainu dialects]. *Minzokugaku kenkyū*, 24(4):307–342.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pretrained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.

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.

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

Hiroshi Nakagawa. 2006. Ainu-jin ni yoru ainu-go hyōki e no torikumi [Ainu people's efforts to write the Ainu language]. *Writing unwritten languages*, pages 1–44.

Hiroshi Nakagawa. 2011. Ainu no shin'yō ni okeru jojutsu-sha no ninshō [The person of the narrator in Ainu mythology]. *Northern Language Studies*, 1:139–156.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2018. A proposal for a unified corpus of the Ainu language. *IPSJ SIG Tech. Rep.*, 237:1–6.

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.

Michal Ptaszynski, K Mukaichi, and Yoshio Momouchi. 2013. Nlp for endangered languages: Morphology analysis, translation support and shallow parsing of Ainu language. In *Proceedings of the 19th Annual Meeting of The Association for Natural Language Processing, Nagoya, Japan*, pages 12–15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Tomomi Sato. 2012. Ainu-go no genjō to fukkō [The state of Ainu language and revitalization]. *Gengo Kenkyū*, 142:29–44.

Suzuko Tamura. 2020. *Ainu-go no sekai [The World of Ainu Language]*, shinsō fukyū-ban edition. Yoshikawa kōbunkan.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Full List of Resources

In this section, we list the resources included in our corpus.

1. **A Glossed Audio Corpus of Ainu Folklore**[4]**:** A resource provided by NINJAL, comprising 30 stories of Ainu oral literature from the Chitose and Saru regions. Each entry includes part-of-speech classification and English translations.

2. **A Topical Dictionary of Conversational Ainu**[5]**:** A practical phrasebook of the Ainu language, originally compiled by Shozaburo Kanazawa in 1898. For this study, we utilized a transcribed version hosted by NINJAL.

3. **Ainu Shin'yōshū (revised by Tatsumine Katayama):** A collection featuring modern Japanese translations of Yukie Chiri's *Ainu Shin'yōshū*, alongside texts transcribed in the modern writing system. It comprises 13 yukar tales. Although Hideo Kirikae also produced a modern revision, we adopted Katayama's edition because it includes symbols indicating the boundaries of personal affixes, which follows *AKOR ITAK* orthography.

4. **Ainu Textbooks by The Foundation for Ainu Culture**[6]**:** A series of textbooks for learning the Ainu language, published by The Foundation for Ainu Culture. Available as PDFs on their website, these materials cover eight dialects across three proficiency levels.

5. **AKOR ITAK:** A textbook published by the Hokkaido Utari Association in 1994. It features the Ainu language in various dialects and explains basic vocabulary, rituals, and folklore. It is also known for proposing an Ainu orthography for the Latin alphabet and Katakana. For this study, we used only the grammar lesson sections containing conversation examples.

6. **Bulletin of the Hokkaido Ainu Culture Research Center**[7]**:** A research bulletin featuring papers with transcriptions in both Ainu and Japanese. For this study, we excerpted articles from issues 9, 10, 11, 12, and 17.

7. **Collection of Ainu Oral Literature**[8]**:** A collection of oral literature from Biratori Town, compiled by Shigeru Kayano and transcribed

[4] https://ainu.ninjal.ac.jp/folklore/

[5] https://ainu.ninjal.ac.jp/topic/dictionary/en/

[6] https://www.ff-ainu.or.jp/web/potal_site/details/post.html

[7] https://ainu-center.hm.pref.hokkaido.lg.jp/05_001.htm

[8] https://nibutani-ainu-museum.com/culture/language/story/

by researchers at Chiba University as part of a research project by the Agency for Cultural Affairs. This collection is available on the website of the Nibutani Ainu Culture Museum.

8. **Dictionary of the Mukawa dialect of Ainu**[9]**:** A phrasebook of the Mukawa dialect of Ainu, compiled by Tatsumine Katayama and available in CSV format on the website of the Graduate School of Humanities and Social Sciences at Chiba University.

9. **ILCAA Online Text of Ainu Collected by Suzuko Tamura**[10]**:** A website featuring online texts of the Ainu language with Japanese translations, collected by Suzuko Tamura. Audio recordings accompany each sentence.

10. **Learning Ainu Language by Listening to Kamuy Yukar:** A textbook by Hiroshi Nakagawa, published by Hakusuisha, focusing on Ainu grammar through Kamuy Yukar from the Chitose region.

11. **New Express Plus Ainu-go:** Another textbook by Hiroshi Nakagawa, published by Hakusuisha, focusing on everyday conversations from the Saru region and explaining Ainu grammar.

12. **The Ainu Language Archive**[11]**:** A website maintained by the National Ainu Museum, providing the largest corpus of Ainu language texts alongside their Japanese translations.

13. **The Ainu Times:** A periodical published by the Ainu Language Pen Club, consisting of essays, news articles, and various other writings. Since its inception in 1997, it has released 80 issues. Articles are contributed by volunteers, with each piece indicating the specific dialect used at the conclusion of the text. In this study, we utilized 71 issues from No. 3 to No. 80.

[9] http://itelmen.placo.net/Ainu-archives/mukawa/

[10] https://online-resources.aa-ken.jp/resources/detail/IOR000018

[11] https://ainugo.nam.go.jp/