Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Ainu–Japanese Bi-directional Neural Machine Translation: A Step Towards Linguistic Preservation of Ainu, An Under-Resourced Indigenous Language in Japan*

So Miyagawa[1]

[1]University of Tsukuba, Japan

Corresponding author: So Miyagawa , miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

This study presents a groundbreaking approach to preserving the Ainu language, recognized as critically endangered by UNESCO, by developing a bi-directional neural machine translation (MT) system between Ainu and Japanese. Utilizing the Marian MT framework, known for its effectiveness with resource-scarce languages, the research aims to overcome the linguistic complexities inherent in Ainu's polysynthetic structure. The paper delineates a comprehensive methodology encompassing data collection from diverse Ainu text sources, meticulous preprocessing, and the deployment of neural MT models, culminating in the achievement of significant SacreBLEU scores that underscore the models' translation accuracy. The findings illustrate the potential of advanced MT technology to facilitate linguistic preservation and educational endeavors, advocating for integrating such technologies in safeguarding endangered languages. This research not only underscores the critical role of MT in bridging language divides but also sets a precedent for employing computational linguistics to preserve cultural and linguistic heritage.

Keywords

Ainu; machine translation; low-resource language; under-resource language; natural language processing; language revitalization; language reclamation; endangered language education

I INTRODUCTION

The Ainu language, known for its complex polysynthetic structure and cultural depth, is traditionally used by the Ainu people who reside in Japan's northern territories, including Hokkaido, Southern Sakhalin, and the Kuril Islands. Despite its complexity, this language is facing a severe threat of extinction. In 2009, UNESCO declared Ainu as "critically endangered" Moseley [2010], highlighting the urgent need for preservation initiatives. The precarious state of the Ainu language is emphasized by the reduced number of native speakers and the disappearance of several dialects, such as those of Sakhalin Ainu and Kuril Ainu.

The Ainu language exhibits unique linguistic features like polysynthesis and noun incorporation, which are also found in numerous indigenous languages across North America and Northeast Siberia. The examples below (1, 2) illustrate these characteristics of polysynthesis and noun incorporation.

(1) Polysynthesis of Hokkaido Ainu [Shibatani, 1990, 72]

Usa-opuspe *a-e-yay-ko-tuyma-si-ram-suy-pa*
various-rumors 1SG-APL1-REFL1-APL2-far-REFL2-heart-sway-ITR

"I wonder about various rumors."[1]

(2) Noun incorporation of Hokkaido Ainu (Kobayashi, 2008, 207, translated and annotated by Miyagawa, 2023, 568)

a. Without noun incorporation

set *a=cari*
cage 1SG=reveal

"I will reveal the cage"

b. With noun incorporation

set-cari=an
cage-reveal=1SG

"I will reveal (a/the) cage"

In this context, the present study is set to harness the latest developments in Natural Language Processing (NLP) and Machine Translation (MT) technologies to deepen our comprehension and translation capabilities concerning the Ainu language. The goal is to utilize these cutting-edge technologies to aid in preserving and revitalizing the Ainu language, especially considering the critical status highlighted by UNESCO.

This research aims to create an AI-powered educational program and a teaching robot designed to support the teaching and safeguarding of the Ainu language. The envisioned program plans to integrate various elements such as speech recognition and production, part-of-speech tagging, and Universal Dependencies tagging, leveraging recent linguistic research on the Ainu language.

Foundational work by researchers like Nowakowski et al. [2019], who developed the Mingmatch—a n-gram model for segmenting Ainu words, and Matsuura et al. [2020b], who compiled an Ainu folklore speech corpus, provides critical underpinnings for this study. Building upon these seminal contributions, this project seeks to construct a comprehensive NLP model that utilizes the Marian MT as its core for translating between Ainu and Japanese. The insights derived from this research will guide the creation of AI-supported educational resources, thus aiding in preserving and understanding the Ainu language and culture.

[1]The list of abbreviations in the gloss: 1SG = first-person singular, APL = applicative, REFL = reflective, ITR = iterative.

II NLP FOR AINU

The exploration of advanced computational methods like NLP and MT in the context of language revival is increasingly gaining traction. The pioneering work of Nowakowski et al. [2019] introduced an efficient n-gram model for the segmentation of Ainu words, demonstrating the capabilities of computational techniques to enhance the study and accessibility of the Ainu language. Following this, Nowakowski [2020] took further strides by developing a digital corpus alongside fundamental language technologies for Ainu. Additionally, Nowakowski et al. [2017] advanced the development of more effective text-processing tools for Ainu, setting a foundation for applying NLP in its digitization.

Subsequent advancements by Nowakowski [Nowakowski et al., 2023] involved refining a multilingual speech representation model for lesser-resourced languages via multilingual fine-tuning and ongoing pretraining, illustrating the adaptability of NLP techniques to languages with limited resources like Ainu.

In parallel, advancements in applying speech recognition technologies to Ainu were realized by Matsuura et al. [2020b], who compiled a speech corpus of Ainu folklore and implemented end-to-end speech recognition. Furthermore, they explored the use of generative adversarial training data adaptation for enhancing automatic speech recognition in scenarios of minimal resources Matsuura et al. [2020a], significantly forwarding the application of NLP in low-resource language contexts.

In the linguistic analysis of Ainu, significant contributions were made by Senuma and Aizawa [2017] and Yasuoka [2021] in developing universal dependencies for Ainu and Ptaszynski et al. [2016] in refining part-of-speech tagging for the language, which are crucial for crafting precise NLP tools.

The broader challenges associated with MT have been critically discussed in the works of Koehn and Knowles [2017], highlighting the necessity for sophisticated techniques to surmount these obstacles effectively. From an educational standpoint, the innovative use of NLP tools was showcased by Nowakowski et al. [2020] through developing an Ainu-speaking Pepper robot, underlining the potential of such technologies in enhancing and safeguarding endangered languages.

These studies contribute valuable insights and methodologies and open avenues for further investigation into the application of MT and NLP technologies for the preservation of endangered languages like Ainu.

III CORPUS BUILDING

Our methodological framework was initiated with a meticulous preprocessing stage, sourcing data from various digital Ainu texts. We tapped into several invaluable resources for this purpose: the Ainugo Archive hosted by the National Ainu Museum[2], offering a rich repository of Ainu language materials; the Glossed Audio Corpus of Ainu Folklore provided by the National Institute for Japanese Language and Linguistics [3], which includes annotated audio recordings; and the ILCAA Ainu Language Resource made available by the Tokyo University of Foreign Studies[4], further enriching our corpus with diverse linguistic inputs. Transcribing texts from

[2] https://ainugo.nam.go.jp/ (accessed June 24, 2023)
[3] https://ainu.ninjal.ac.jp/folklore/ (accessed June 24, 2023)
[4] https://ainugo.aa-ken.jp/ (accessed June 24, 2023)

Katakana to the Roman alphabet was a pivotal initial step, setting the stage for the subsequent refinement process aimed at purging the corpus of duplicative or inconsistent data entries. This phase was critical to ensure the integrity and uniformity of the dataset before tokenization and segmentation into sentence pairs. Given the intrinsic polysynthetic characteristics of the Ainu language, which involves the complex intertwining of morphemes to form words, our tokenization efforts were especially cautious and deliberate to avoid misinterpretation or loss of linguistic nuances, as exemplified in Examples (1) and (2). The resulting dataset comprised around 100,000 sentence pairs, marking a significant corpus for our translation model training.

3.1 Ainu orthography

The Ainu language, indigenous to the northern regions of Japan, including Hokkaido and parts of Sakhalin and the Kuril Islands in Russia, has a rich oral tradition but a relatively recent history of being written down. Developing a writing system for Ainu has involved several scripts, including Kanji, Hiragana, Katakana, the Roman alphabet, and even the Cyrillic alphabet. Each script has been employed under different circumstances and for various purposes, reflecting the Ainu people's complex history and cultural exchanges.

3.2 Kanji

Historically, the Ainu language did not have a writing system until the modernization. When the Ainu language began to be documented by Japanese scholars, Kanji (Chinese characters used in Japanese writing) was occasionally used to represent Ainu words. However, given the significant linguistic differences between Japanese and Ainu, Kanji was not suitable for accurately capturing the sounds of the Ainu language. Instead, Kanji was primarily used in a limited and symbolic manner, often for names or specific terms within Japanese texts that referred to Ainu concepts or culture.

3.3 Hiragana

Hiragana, a syllabic script used in the Japanese language, was one of the first systems adopted for writing Ainu. Hiragana was used to approximate the phonetic sounds of Ainu words. This script was more adaptable than Kanji for transcribing Ainu sounds, although it was still an imperfect match due to the differences in phonetics between the two languages. Hiragana was primarily used by Japanese researchers and educators, especially in the early stages of Ainu language documentation and study.

3.4 Katakana

Katakana, another syllabic script used in Japanese, became the preferred system for writing Ainu among Japanese scholars by the late 19th and early 20th centuries. Katakana was considered more suitable for representing the sounds of non-Japanese words, including Ainu. It has been used extensively in academic works, language instruction materials, and dictionaries dedicated to the Ainu language. Today, Katakana remains one of Japan's most common scripts for writing Ainu, especially in educational contexts and scholarly research.

3.5 Cyrillic Alphabet

The use of the Cyrillic alphabet for Ainu is far less common and primarily restricted to regions under Russian influence, such as Sakhalin. The Cyrillic script has been used sporadically for Ainu by Russian researchers and during periods when Sakhalin has been under Russian control or influence. However, the Cyrillic alphabet's use for Ainu is limited compared to the other scripts mentioned.

3.6 Roman Alphabet

The Roman alphabet (or Latin script) was introduced for writing Ainu primarily through the work of foreign and Japanese linguists and missionaries in the 19th and 20th centuries. It offered a higher degree of phonetic accuracy for representing Ainu sounds. It has been used in various linguistic studies such as Sato [2008], textbooks, and translations of Ainu folklore and oral literature. Using the Roman alphabet has lowered the threshold of the international research of Ainu and increased its accessibility to non-Japanese speakers. Many contemporary Ainu language revitalization efforts utilize the Roman alphabet for teaching materials and documentation.

The history of writing the Ainu language reflects a tapestry of cultural interactions and scholarly efforts to document and preserve the language. From the symbolic use of Kanji to the phonetic adaptations of Hiragana and Katakana and the broader phonetic capabilities of the Roman alphabet, each script has played a role in the ongoing journey of Ainu language preservation. The script choice often depends on the context, purpose, and audience for the Ainu language material, with Katakana and the Roman alphabet being the most prevalent in contemporary times.

Our methodological framework was initiated with a meticulous preprocessing stage, sourcing data from various digital Ainu texts. We tapped into several invaluable resources for this purpose: the Ainugo Archive hosted by the National Ainu Museum[5], offering a rich repository of Ainu language materials; the Glossed Audio Corpus of Ainu Folklore provided by the National Institute for Japanese Language and Linguistics [6], which includes annotated audio recordings; and the ILCAA Ainu Language Resource made available by the Tokyo University of Foreign Studies[7], further enriching our corpus with diverse linguistic inputs. Transcribing texts from Katakana to the Roman alphabet was a pivotal initial step, setting the stage for the subsequent refinement process aimed at purging the corpus of duplicative or inconsistent data entries. This phase was critical to ensure the integrity and uniformity of the dataset before tokenization and segmentation into sentence pairs. Given the intrinsic polysynthetic characteristics of the Ainu language, which involves the complex intertwining of morphemes to form words, our tokenization efforts were especially cautious and deliberate to avoid misinterpretation or loss of linguistic nuances, as exemplified in Examples (1) and (2). The resulting dataset comprised around 100,000 sentence pairs, marking a significant corpus for our translation model training.

IV MACHINE LEARNING

In our comprehensive study, we strategically harnessed the Marian MT framework, renowned for its exceptional efficiency and adaptability as an open-source machine translation tool.

The foundational framework of this application is following the example of Coptic Translator [Enis and Megalaa, 2023]. This application utilizes a method that employs a multilingual model based on the Transformer architecture. It is available via the Hugging Face library and has been specifically trained with Coptic-English translation data. The project takes advantage of models trained based on Marian MT [Junczys-Dowmunt et al., 2018], offered by Hugging Face, including the 'opus-mt-mul-en' and 'opus-mt-en-mul' machine translation models. These were developed by the Helsinki-NLP team at the University of Helsinki, leveraging the OPUS

[5] https://ainugo.nam.go.jp/ (accessed June 24, 2023)
[6] https://ainu.ninjal.ac.jp/folklore/ (accessed June 24, 2023)
[7] https://ainugo.aa-ken.jp/ (accessed June 24, 2023)

corpus for training [Tiedemann and Thottingal, 2020]. These models are integrated into a comprehensive framework that supports machine translation across 119 languages. The developer has effectively created and evaluated bidirectional translation models for Japanese-Ainu and Ainu-Japanese, utilizing these foundational models.

This choice was primarily due to Marian MT's proven prowess in handling projects involving languages with scarce resources, as evidenced by previous research Ponti et al. [2021]. Its capability to adeptly manage various linguistic structures made it an optimal choice for our work with the Ainu language. This polysynthetic language presents unique challenges in translation due to its complex morphological constructions Ortega et al. [2020].

The training of the Marian MT model was conducted with this carefully prepared corpus, encompassing translations from Ainu to Japanese and vice versa. Model optimization was a critical aspect of this process, utilizing a learning rate schedule to dynamically adjust the learning rate to preclude the model from overfitting to the training data. This approach, complemented by the implementation of early stopping, ensured that the training ceased at the optimal moment when the model's performance on a validation set plateaued, thereby preventing the model from memorizing the training data at the expense of its ability to generalize Almansor and Al-Ani [2018].

To evaluate our machine translation system, we employed the SacreBLEU metric Kim and Kim [2022b], a robust and reliable standard for measuring translation quality. SacreBLEU's consistent approach to tokenization and detokenization across evaluations is particularly advantageous, ensuring that comparisons between different MT systems or iterations are fair and uniform. Furthermore, its capability to handle multiple reference translations is paramount for languages like Ainu, where the scarcity of parallel texts means that a single sentence may have several acceptable translations. This comprehensive evaluation method allows for a nuanced assessment of translation accuracy, which is critical in the context of language preservation efforts Kim and Kim [2022a].

Our methodology has established a robust and highly accurate machine translation system by integrating the Marian MT framework with a thorough preprocessing of the Ainu language corpus, meticulous training in both translation directions, and a detailed evaluation using the SacreBLEU metric. This system not only facilitates translations between Ainu and Japanese but also represents a significant step forward in the preservation and accessibility of the Ainu language, contributing to broader efforts to safeguard this valuable cultural heritage.

V RESULTS

This segment elucidates the findings from our comprehensive machine translation (MT) experiments, offering an in-depth analysis of the outcomes and their broader implications. Central to our investigation were the translation endeavors from Japanese to Ainu and from Ainu to Japanese, with a particular focus on evaluating the efficacy of translations in these two directions[8] (See Table 1).

In the Japanese-to-Ainu translation task, the model was fine-tuned on a vast dataset aggregated

[8]The models crafted during this research are accessible on HuggingFace, providing resources for Ainu-to-Japanese translation: https://huggingface.co/SoMiyagawa/ainu-2-japanese, Japanese-to-Ainu translation: https://huggingface.co/SoMiyagawa/japanese2ainu, and for bi-directional translation efforts: https://huggingface.co/SoMiyagawa/AinuTrans-2.0 (all accessed on June 24, 2023).

	Jpn.-Ain.	Ain.-Jpn.	Bi-dir.
Num. pairs	97,161	95,232	220,023
SacreBLEU	32.90	10.45	29.91

Table 1: Dataset sizes and SacreBLEU scores for the top-performing MT models in each translation scenario

from diverse Ainu digital text repositories. Achieving a SacreBLEU score of 32.90, the model demonstrated substantial accuracy, illustrating its proficiency in navigating the complex linguistic terrain between these two languages. This level of precision in translation is particularly noteworthy, considering the significant challenges inherent in devising an MT system tailored to a language with as scarce resources as Ainu.

Transitioning to the task of translating Ainu to Japanese, we encountered additional obstacles, predominantly stemming from the scarcity of comprehensive resources in the Ainu language. Despite these limitations, our MT system attained a SacreBLEU score of 10.45, underscoring its operational effectiveness even in less-than-ideal conditions. Furthermore, we ventured into experiments with Marian MT on a bidirectional basis, employing an enriched corpus that was meticulously reversed for the second phase of the dataset. This bidirectional approach yielded a SacreBLEU score of 29.91, affirming the model's adeptness at facilitating translations irrespective of the direction, thereby enabling seamless communication across languages.

The positive SacreBLEU scores garnered from our experiments highlight MT's potential utility in supporting language preservation and revitalization projects. The data indicates that even within the constraints of limited linguistic resources, MT models can achieve a degree of accuracy that renders them valuable tools for learners and scholars engaged with the Ainu language Kim and Kim [2022b].

Moreover, our findings resonate with and augment the accomplishments of earlier endeavors that applied computational techniques to the Ainu language. A prime example is the pioneering speech recognition initiative conducted by Kawahara Lab at Kyoto University, with their findings detailed in Matsuura et al. [2020b]. Together, these collective efforts underscore the transformative potential of NLP and MT technologies in preserving and revitalizing endangered languages, offering new avenues for academic and cultural engagement.

These experimental outcomes validate the feasibility of MT as a means to facilitate language learning and cultural preservation and signal the advent of a new era in linguistic studies. They advocate for a future where MT and computational linguistics are pivotal in bridging linguistic divides, enhancing accessibility to cultural narratives, and empowering communities by preserving their linguistic heritage.

Looking forward, the path is set for an array of explorations into the realm of MT applications within contexts of under-resourced languages. Future research endeavors could pivot towards refining the accuracy and efficiency of these models, expanding upon the datasets to encompass a broader linguistic spectrum, and delving into how these technological advancements can be seamlessly integrated into interactive and immersive language learning platforms. Such initiatives promise to further the cause of Ainu language and culture revitalization and establish a template for preserving other endangered languages, thereby enriching our global linguistic and cultural mosaic.

VI CONCLUSIONS

This study explored the utilization of machine translation (MT) for safeguarding and revitalizing the Ainu language, identified by UNESCO as critically endangered and characterized by its limited resources. Employing the neural MT framework, specifically Marian MT, and leveraging an extensive collection of data from varied Ainu digital text sources, our research has substantially demonstrated MT's practicality and effectiveness in preserving languages.

The achievements of this research are notable. Achieving a SacreBLEU score of 32.90 in the task of translating from Japanese to Ainu highlights the high quality of translations despite the Ainu language's complex polysynthetic structure Ortega et al. [2020]. Furthermore, the accomplishment of a SacreBLEU score of 29.91 in the bi-directional translation task between Japanese and Ainu attests to the capability of the neural MT framework to handle intricate and resource-scarce languages efficiently Kim and Kim [2022b].

Our findings add to the growing literature on the potential of MT to overcome linguistic barriers and support the conservation of endangered languages. This study is in line with other research, like the work of Ranathunga et al. [2023] on the application of neural MT to low-resource languages, and Kumar et al. [2021], which investigated the use of MT in the context of languages with scant resources.

When evaluating machine translations for agglutinative languages like Ainu and Japanese, the character F-score (chrF; Popović, 2015) is particularly suitable due to its focus on character-level accuracy, capturing the nuances of morphological details. ChrF assesses the precision and recall of characters in the translated text, making it adept at evaluating languages where morphological complexity is a defining feature. This metric is beneficial for reflecting the quality of translations that handle the inflection and concatenation of morphemes, which are common in agglutinative languages.

However, relying solely on chrF might not offer a complete evaluation. The BLEU [Papineni et al., 2002] or SacreBLEU score, which measures the match of n-grams between the translated and reference texts, provides insights into the fluency and adequacy of translations at a broader linguistic level. Although BLEU / SacreBLEU might not fully capture the morphological richness of agglutinative languages, it remains a standard for assessing translation quality, focusing on phrase matching and word order.

Utilizing both chrF and BLEU / SacreBLEU offers a more comprehensive evaluation by combining morphological accuracy (chrF) with overall fluency and adequacy (BLEU/SacreBLEU). This dual approach allows for a balanced assessment of machine translation performance, acknowledging the complexity of translating agglutinative languages while also considering general translation quality standards.

The study highlights the imperative for ongoing advancement in MT approaches for languages with minimal resources, especially when conventional linguistic resources are sparse or absent. By merging insights from Natural Language Processing (NLP), MT, and language conservation efforts, our research provides a model for future studies and underscores the critical need for innovation in these domains Pilch et al. [2022].

The endeavor to revive endangered languages involves a broad spectrum of activities requiring the collaboration of linguists, educators, technologists, and native communities. Our findings affirm that MT and related linguistic technologies are indispensable in this intricate process. Although further model refinements and dataset expansions are necessary to improve translation

quality, our present results highlight MT's vital contribution to language preservation efforts.

In summary, our investigation suggests that MT can significantly enhance the accessibility of languages with limited resources, such as Ainu. By enabling communication, safeguarding cultural heritage, and deepening the appreciation for diverse human narratives, our study validates the critical importance of language conservation and the role of technology in facilitating these objectives. Given the critically endangered status of Ainu as determined by UNESCO, our research contributes to the battle against linguistic extinction, underscoring the value of preserving our collective linguistic legacy Moseley [2010].

References

Ebtesam H Almansor and Ahmed Al-Ani. A hybrid neural machine translation technique for translating low resource languages. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part II 14*, pages 347–356. Springer, 2018.

Maxim Enis and Andrew Megalaa. Ancient voices, modern technology: Low-resource neural machine translation for coptic texts, 2023. Coptic Translator, https://www.coptictranslator.com/paper.pdf (2024211).

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. Marian: Fast neural machine translation in C++. In Fei Liu and Thamar Solorio, editors, *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4020.

Ahrii Kim and Jinhyeon Kim. Vacillating human correlation of sacrebleu in unprotected languages. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 1–15, 2022a.

Ahrii Kim and Jinhyun Kim. Guidance to Pre-tokeniztion for SacreBLEU: Meta-Evaluation in Korean. 2022b.

Miki Kobayashi. Ainu-go no meishi hougou [noun incorporation in ainu]. *Journal of Studies on Humanities and Public Affairs of Chiba University*, 17:199–214, 2008.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*, 2021.

Kohei Matsuura, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. Generative adversarial training data adaptation for very low-resource automatic speech recognition. *arXiv preprint arXiv:2005.09256*, 2020a.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. *arXiv preprint arXiv:2002.06675*, 2020b.

So Miyagawa. Noun incorporation in coptic. In Heike Sternberg-el Hotabi Diliana Atanassova, Frank Feder, editor, *Pharaonen, Mönche und Gelehrte: Auf dem Pilgerweg durch 5000 Jahre ägyptische Geschichte über drei Kontinente. Heike Behlmer zum 65. Geburtstag*, volume 4 of *Texte und Studien zur Koptischen Bibel*. Harassowitz Verlag, Wiesbaden, 2023.

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

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. Towards better text processing tools for the Ainu language. In *Language and Technology Conference*, pages 131–145. Springer, 2017.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. Mingmatch—a fast n-gram model for word segmentation of the Ainu language. *Information*, 10(10):317, 2019.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. Spicing up the game for underresourced language learning: Preliminary experiments with Ainu language-speaking Pepper robot. In *The 6st workshop on linguistic and cognitive approaches to dialog agents*, 2020.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148, 2023.

Karol Piotr Nowakowski. Development of a digital corpus and core language technologies for the Ainu language. 2020.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of

machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Agnieszka Pilch, Ryszard Zygała, and Wiesława Gryncewicz. Quality assessment of translators using deep neural networks for polish-english and english-polish translation. In *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 227–230. IEEE, 2022.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. Modelling latent translations for cross-lingual transfer. *arXiv preprint arXiv:2107.11353*, 2021.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Michal Ptaszynski, Karol Nowakowski, Yoshio Momouchi, and Fumito Masui. Comparing multiple dictionaries to improve part-of-speech tagging of Ainu language. In *Proceedings of the 22nd Annual Meeting of The Association for Natural Language Processing, Sendai, Japan*, pages 7–11, 2016.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

Tomomi Sato. *Ainugo Bumpo no Kiso [Ainu Grammar Basics]*. Daigaku Shorin, Tokyo, 2008.

Hajime Senuma and Akiko Aizawa. Toward universal dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 133–139, 2017.

Masayoshi Shibatani. *The languages of Japan*. Cambridge University Press, 1990.

Jörg Tiedemann and Santhosh Thottingal. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 2020.

Koichi Yasuoka. Universal Dependencies ni yoru Ainugo tekisuto koopasu [Ainu text corpus by Universal Dependencies] . *IPSJ SIG Technical Report*, 2021-CH-127:1–8, 2021.