Mergen: The First Manchu-Korean Machine Translation Model Trained on Augmented Data

Jean Seo, Sungjoo Byun, Minha Kang, Sangah Lee
Seoul National University
{seemdog, byunsj, alsgk1123, sanalee}@snu.ac.kr

Abstract

The Manchu language, with its roots in the historical Manchurian region of Northeast China, is now facing a critical threat of extinction, as there are very few speakers left. In our efforts to safeguard the Manchu language, we introduce *Mergen*, the first-ever attempt at a Manchu-Korean Machine Translation (MT) model. To develop this model, we utilize valuable resources such as the Mǎnwén Lǎodàng(a historical book) and a Manchu-Korean dictionary. Due to the scarcity of a Manchu-Korean parallel dataset, we expand our data by employing word replacement guided by GloVe embeddings, trained on both monolingual and parallel texts. Our approach is built around an encoder-decoder neural machine translation model, incorporating a bi-directional Gated Recurrent Unit (GRU) layer. The experiments have yielded promising results, showcasing a significant enhancement in Manchu-Korean translation, with a remarkable 20-30 point increase in the BLEU score.

1 Introduction

Efforts to conserve and revive endangered languages have surged, with modern advancements in Natural Language Processing (NLP) playing a pivotal role. Zhang et al. (2020) introduce ChrEn, a Cherokee-English parallel dataset, and examine methodologies like Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). Zhang et al. (2020) aid the conservation of Cherokee, a critically endangered Native American dialect. On a similar note, Luo et al. (2020) present a decipherment model for lost languages that addresses challenges posed by non-segmented scripts and undetermined proximate languages, leveraging linguistic constraints and the International Phonetic Alphabet (IPA) for phonological patterns.

Manchu language, originated from the historical Manchurian region in Northeast China, stands as a highly endangered Tungusic language of East Asia (Tsunoda, 2006). There are merely few Manchu speakers left nowadays, leading Manchu to be labeled 'nearly extinct' by UNESCO (Kim et al., 2008). The Manchu spell checker (You, 2014) and the Manchu corpus with morphological annotations (Choi et al., 2023a,b) are the only prior approaches to embrace Manchu in the field of NLP. We introduce *Mergen*, the first Manchu-Korean machine translation model, which marks the pioneering effort to apply MT to the Manchu language.

We employ two sets of parallel corpora for machine translation from Manchu to Korean, as detailed in Kim et al. (2019). Initially, we train an adapted version of the NMT model (Bahdanau et al., 2016). Assuming the unexpectedly low performance is due to the scarcity of Manchu-Korean data, we augment the size of parallel data several fold utilizing GloVe (Pennington et al., 2014). Our findings suggest that this data augmentation methodology substantially enhances translation quality.

Despite the constrained availability of resources, our goal is to enhance Manchu-Korean machine translation performance. To symbolize our commitment to the field of Manchu NLP, we christen our model *Mergen*, denoting a sage or a wise individual in the Manchu lexicon. Our translation approach, which employs a data augmentation technique, not only seeks to improve Manchu-Korean translation performance but also aims to eventually serve as a potential model for addressing NLP challenges in other extremely low-resource scenarios as addressed in King (2015).

2 Related Work

2.1 Low-Resource Machine Translation

MT necessitates parallel data of source and target languages to be trained effectively. However, the majority of language pairs face a scarcity of resources. As a result, there has been various research

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 118–124
December 7, 2023 ©2023 Association for Computational Linguistics

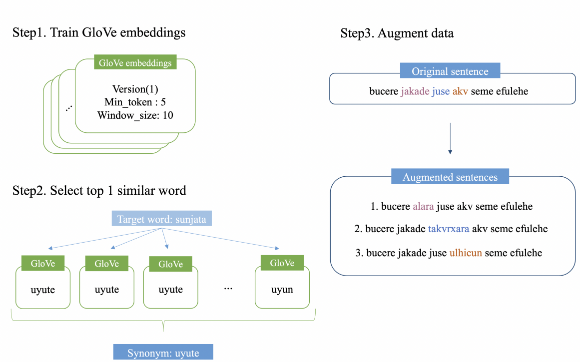

Figure 1: Our data augmentation methodology. First, we train ten versions of GloVe embedding models, varying in the minimum token length of source data and window size. Then, the presumable synonym for the target word is selected via comparing the frequency of outputs from each model. Finally, we augment data through replacing original words with synonyms if possible. The pair of original and substituted words are in the same color.

endeavors aimed at developing translation models in low-resource scenarios. Extended language models such as XLM-RoBERTa (Conneau et al., 2019), mBART (Tang et al., 2021), multilingual BERT (mBERT) (Pires et al., 2019), and mT5 (Xue et al., 2021) are trained on diverse languages. Yet, most of these multilingual language models tend not to incorporate endangered languages. This leads to an increasing disparity in NLP resources, where less-resourced languages are further marginalized. Numerous strategies have been attempted in low-resource machine translation. Gibadullin et al. (2019) and Siddhant et al. (2020) employ monolingual data in low-resource NMT. Additionally, utilization of pre-trained word embeddings (Qi et al., 2018) and application of transfer learning with pre-trained language models like XLM (Lample and Conneau, 2019) and mBART (Liu et al., 2020) have been employed. Furthermore, Lakew et al. (2018) enhance the zero-shot translation capability of low-resource languages.

2.2 Typological Similarities between Manchu and Korean

There are several typological motivations for translating Manchu to Korean using a Machine Translation model. The genetic affinity between Manchu and Korean is not proven, but it is well-known that Manchu has a similar structure to that of Korean. The word order of Manchu and Korean mostly coincide, including the order of 'noun-particle,' 'modifier-modified,' and 'object-verb,' etc. (Park, 2018). Substitutes in Korean, *kes*, and Manchu, *-ngge*, have analogous grammatical functions and positions (Choi, 2009). The two languages both show factivity alternation by using the attitude verb 'to know' (Lee, 2019) and have parallel subordinated clause structures (Malchukov and Czerwinski, 2020). These typological similarities between Manchu and Korean arouse interest in understanding and linguistically translating each other. In fact, studies of the Manchu language are active in Korea (Ko, 2023).

3 Data

3.1 Materials

The Manchu corpora used in this study comprise all of the digitized textual data available and can be categorized as either parallel or monolingual. The parallel corpora are *Mănwén Lăodàng* (1774-1778) and the Manchu-Korean dictionary. These corpora consist of Manchu texts and their corresponding translations in Korean. We only utilize a section of the *Mănwén Lăodàng* and its translations from Kim et al. (2019), which details the history of Nurhaci, the Emperor Taizu of Qing dynasty. Additionally, we refer to the dictionary from Lee (2017) and select sentences with a minimum of three words.

The monolingual texts of Manchu include the remaining part of *Mănwén Lăodàng*, Manchu-Manchu dictionaries, and several pieces of literature. The part of *Mănwén Lăodàng* left over is the chronicle of Hong Taiji, the Emperor Taizong of Qing. The Manchu-Manchu dictionaries we use are *Yùzhì Qīngwénjiàn* (1708) and *Yùzhì Zēngdìng Qīngwénjiàn* (c.1771).

The other data is composed of novels, *Ilan gurun i bithe* (c.1723-1735) and *Gin ping mei bithe* (1708). *Ilan gurun i bithe* is the translated version of *The Romance of the Three Kingdoms*. *Gin ping mei bithe* is translated from the Chinese naturalistic novel, *The Plum in the Golden Vase*. The size

Monolingual data	Number of sentences
Mănwén Lăodàng–Taizong	2,220
Ilan gurun i bithe	41,904
Gin ping mei bithe	21,376
Yùzhì Qīngwénjiàn	11,954
Yùzhì Zēngdìng Qīngwénjiàn	18,420
Parallel data (Man-Kor)	
Mănwén Lăodàng–Taizu	22,578
Manchu-Korean Dictionary	40,583

Table 1: The size of each material

description of each data can be found in Table 1.

3.2 Romanization of Manchu script and Hangul

To create a more sufficient translation model, the script of each language should be unified in one writing system. That is, both the source and target language should undergo transliteration to the Latin alphabet, so-called 'romanization'. For the romanization of Manchu, we apply Abkai Latin transliteration. The Abkai romanization suggested by An (1993) is a Pinyin-based writing system. We also use the system of Seong (1977) for the special characters in the Manchu script. Transliteration of Manchu to the Latin alphabet is reversible except for a couple of letters. For the Latin transliteration of Korean, we employ Yale romanization system (Martin, 1992) and develop the corresponding Python library[1]. See Appendix A for examples.

3.3 Data Augmentation

The lack of available Manchu linguistic data poses challenges not only for the pre-training of transformer-based models but also for the training of simpler and more lightweight models, such as encoder-decoder models. Inspired by TinyBERT (Jiao et al., 2020), we adopt a novel data augmentation approach. While the data augmentation method in TinyBERT (Jiao et al., 2020) combines both BERT (Devlin et al., 2019) and GloVe (Pennington et al., 2014), we exclusively employ GloVe embeddings. This decision stems from the absence of a pre-trained BERT model tailored to Manchu and the significant difficulty of pre-training a BERT model from scratch due to the limited amount of available textual data.

Our methodology involves training GloVe embedding models with two different versions of the dataset: (1) a dataset comprising sentences with at least 3 words, and (2) a dataset comprising sentences with at least 5 words. The dataset includes both monolingual and parallel text data. Various window sizes, specifically 1, 3, 5, 7, and 10, are used during the training process, resulting in a total of 10 distinct variations of GloVe embeddings.

For each word in the training dataset, we gather the most similar word predicted by each individual GloVe embedding. Amongst the list of 10 words generated from these separate models, the word with the highest frequency is considered the most suitable synonym for the target word. Following this, we substitute a single word in each sentence from parallel text data with the identified synonym. The augmentation steps are described in Figure 1. This procedure leads to the creation of two augmented versions of the original dataset: full augmentation and half augmentation. The first version involves replacing every word possible in each sentence with its corresponding synonym, significantly expanding the dataset size relative to the average sentence length. The second version is generated by replacing half of the words in each sentence with their respective synonyms, resulting in a dataset expansion about half the size of the first method. Additional details regarding the original and augmented dataset are available in Table 2.

augmentation	Mǎnwén Lǎodàng –Taizu (train)	Man-Kor Dict
Before augmentation	20,320	40,583
Full augmentation	179,843	154,404
Half augmentation	99,506	100,694

Table 2: The number of sentences of parallel text data before and after augmentation

4 Experiments

4.1 Task Details

In the experiment, we merge *Mǎnwén Lǎodàng* with Manchu-Korean dictionary and shuffle them together. The combined dataset is then divided into training, validation, and testing subsets. These subsets are split in an 8:1:1 ratio. In the augmentation process, we first shuffle and then augment the data to even out the word distributions, finally splitting into subsets.

4.2 Model

We adopt the sequence-to-sequence (seq2seq) framework, a deep learning approach designed to transform one sequence into another. Our model is based on the encoder-decoder structure of the NMT (Bahdanau et al., 2016), implemented with bidirectional Gated Recurrent Unit (GRU) layer (Cho et al., 2014). We incorporate two techniques to enhance the performance: packed padded sequences and masking. Packed padded sequences ensure that the RNN processes only the genuine elements of the input sentence, excluding the padded ones. Masking directs the model to deliberately overlook specific components, like attention weights assigned to padded sections.

[1] anonymous author github

Train	Test	BLEU	PPL
Before augmentation (No augmentation)			
Mănwén Lăodàng	Mănwén Lăodàng	0.0	72.50
Man-Kor Dict	Man-Kor Dict	0.0	59.34
Combined	Mănwén Lăodàng	0.0	61.83
	Man-Kor Dict	0.0	61.16
	Combined	0.0	69.62
Half augmentation			
Mănwén Lăodàng	Mănwén Lăodàng	38.38	147.07
Man-Kor Dict	Man-Kor Dict	0.0	174.94
Combined	Mănwén Lăodàng	36.05	192.95
	Man-Kor Dict	**2.37**	36.14
	Combined	27.59	29.22
Full augmentation			
Mănwén Lăodàng	Mănwén Lăodàng	**38.95**	1549.40
Man-Kor Dict	Man-Kor Dict	0.0	158.25
Combined	Mănwén Lăodàng	**37.17**	447.59
	Man-Kor Dict	2.26	46.54
	Combined	**28.00**	41.97

Table 3: Manchu-Korean Translation Performance

4.3 Results and Discussions

We perform machine translation and evaluate the performance on all the available combinations of parallel corpora: *Mănwén Lăodàng*, Manchu-Korean dictionary, and the combined dataset. In particular, we augment the training sets of each corpus to alleviate the data scarcity problem. Table 3 shows the performance of our Manchu-Korean translation models, with BLEU score (Papineni et al., 2002) and Perplexity (PPL) as the metrices. We train each model for 5 epochs and report the one with the best performance.

The first block of Table 3 shows the translation performance based on the original Manchu-Korean parallel corpora. All the experiments here show BLEU scores of 0.0, which represent that none of the test sentences are accurately translated. Most of the predicted translations include the special symbol '<UNK>' instead of proper Korean tokens, possibly due to the small dataset and vocabulary size.

The second block shows the experiment results from the augmented version of the parallel corpora, where up to 50% of the tokens in each sentence are replaced for data augmentation. The third block displays experiments on another augmented version where all tokens with substitutes are replaced. The augmentation procedure increases the size of the training set, resulting in a significant rise in the translation performance. BLEU scores exceed 38 on the *Mănwén Lăodàng* test set, and around 28 on the combined test set. The two versions of the augmented dataset show comparable performance, but replacing all the possible words in the corpus resulted in slightly higher BLEU scores.

Due to data augmentation, the vocabulary for each model is expanded; for example, the original *Mănwén Lăodàng* vocabulary includes 4,335 words, while the full-augmented dataset constructs an expanded vocabulary with 11,089 words. A larger vocabulary and training set may have helped the language model's representation and result in better translation performance. Additionally, most newly induced words are from the augmentation sources which include monolingual Manchu texts, different from our parallel corpora. This expansion of word diversity may have also affected the models' perplexity to increase when they predicted the next words in each sentence.

On the other hand, results on the Manchu-Korean dictionary are consistently very low, and this may have influenced the lower performance of the combined test set. We suppose that it is because the corpus is a dictionary, where each line is a unique word or phrase. The training set and the test set would have much fewer overlaps in their vocabularies, and this could cause a number of '<UNK>' generations in the model prediction.

5 Conclusion

In our exploration of the critically endangered Manchu language, we have made significant strides towards development of low-resource NLP through the development of the Manchu-Korean MT system, "Mergen." Our endeavor to train this model, despite the challenges posed by the scarcity of a Manchu-Korean parallel dataset, demonstrates the potential of an innovative data augmentation strategy. This attempt is also significant in that we have collected all the digitized Manchu text data. By leveraging resources such as "Mănwén Làodăng" and a Manchu-Korean dictionary, and by adopting a word substitution techniqus guided by GloVe embeddings, we have not only built a functional MT system but have also considerably enhanced its accuracy, as evidenced by the increase in the BLEU score. Our encoder-decoder NMT model, equipped with a bi-directional GRU layer, has shown promising results, offering hope for the preservation and accessibility of the Manchu language to future generations. We anticipate that this research will serve as a foundation for further innovations in the realm of endangered language preservation.

Limitations

The main limitation of this study is the scarcity of resources. Numerous Manchu literatures exist in East Asia (Vovin, 2023), including China (Elliott, 2001), Korea (Ko and You, 2012), and Mongolia (Choi, 2014). However, most of them lack an electronic version. The only publicly available Manchu language database is the Manchu Dictionary and Literature DB, created by Seoul National University and supported by the National Research Foundation of Korea.[2] Furthermore, the majority of these resources have not been translated into Korean. To address this gap, we intend to provide supplementary parallel texts translated into Korean for further study. In addition, we plan to implement a cutting-edge method of Transformer-based language model including Manchu language. Knowledge Distillation could be a way for modeling endangered languages, training a small student model based on those languages and improving it with a teacher model based on high-resource languages (Heffernan et al., 2022).

Ethics Statement

The Manchu language, classified as critically endangered, remains underrepresented due to its scarce resources. As such, it has yet to be incorporated into any multilingual language models. This study pioneers Manchu translation efforts, an endeavor previously uncharted. Our primary research objective as NLP practitioners is to prevent the extinction of Manchu language and ensure its preservation. We have no intention of commercializing the translation model. Instead, by making the model publicly available, we aim to facilitate and encourage as many individuals as possible to learn Manchu using our translator. We are committed to continuous collaboration with Manchu language researchers. We endeavor to enhance the performance of our translator and regularly update it with new Manchu data to ensure its accuracy.

References

Shuangcheng An. 1993. *Man Han Da Ci Dian*. Liaoning Minzhu Chubanshe, Shenyang.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

[2]NFR-2012S1A5B4A01035397, available at `http://ffr.krm.or.kr/base/td037/intro_db.html`

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Donggeun Choi. 2014. A study of Manchu language literatures in Mongolia. *The journal of humanities*, 25:275–303.

Dongguen Choi. 2009. A comparative study of substitute - Korean *keos*, Mongolian *yum*, Manchu *-ngge* -. *Mongolian Studies*, 27:205–228.

Woonho Choi, Sunghoon Jung, and Jeongup Do. 2023a. Construction of the Manchu corpus: focusing on *Manwen laodang Taidzu*. *Altai Hakpo*, 33:67–87.

Woonho Choi, Sunghoon Jung, and Jeongup Do. 2023b. Word embeddings for *Manwen laodang* corpus with focus on names of countries and articles. In *Proceedings of the 16th Seoul International Altaistic Conference*, pages 189–204. The Altaic Society of Korea.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mark C Elliott. 2001. The manchu-language archives of the qing dynasty and the origins of the palace memorial system. *Late Imperial China*, 22(1):1–70.

Ilshat Gibadullin, Aidar Valeev, Albina Khusainova, and Adil Khan. 2019. A survey of methods to leverage monolingual data in low-resource neural machine translation. *CoRR*, abs/1910.00373.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Juwon Kim, Dongho Ko, Gyeyeong Choe, Sangchul Park, Jeongup Do, Hyungmi Lee, Hui Jin, and Jaehong Shim. 2019. *Tongki Fuka Sindaha Hergen i Dangse - The Chronicles of Early Qing Dynasty: Taizu Vol. 1 2*. Seoul National University Press, Seoul.

Juwon Kim, Dongho Ko, Youfeng Han, Lianyu Piao, and B. V. Boldyrev. 2008. *Materials of spoken Manchu*. unesco.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages*. Ph.D. thesis.

Dongho Ko. 2023. Manchu-tungus studies in korea: Focusing on the studies of third-generation scholars. *Reosiahag*, 26:1–27.

Dongho Ko and Hyunjo You. 2012. For building a database of written Manchu. *Kenci Inmwunhak*, 8:5–30.

Surafel Melaku Lakew, Quintino F. Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Improving zero-shot translation of low-resource languages. *CoRR*, abs/1811.01389.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Chungmin Lee. 2019. Factivity alternation of attitude 'know' in Korean, Mongolian, Uyghur, Manchu, Azeri, etc. and content clausal nominals. *Journal of Cognitive Science*, 20(4):449–503.

Hoon Lee. 2017. *Manju Solho Gisun Kamcibuha Buleku Bithe - Manhan-sacen*. Korea University Press, Seoul.

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

Jiaming Luo, Frederik Hartmann, Enrico Santus, Yuan Cao, and Regina Barzilay. 2020. Deciphering undersegmented ancient scripts using phonetic prior. *CoRR*, abs/2010.11054.

Andrej Malchukov and Patryk Czerwinski. 2020. Complex constructions in the Transeurasian languages. In Martine Robbeets and Alexander Savelyev, editors, *The Oxford Guide to the Transeurasian Languages*, pages 625–644. Oxford University Press, Oxford.

Samuel E. Martin. 1992. *A Reference Grammar of Korean*. Charles E. Tuttle, Rutland, VT and Tokyo.

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.

Sangchul Park. 2018. The function of Modern Korean as in discourse. *Eoneohag*, 81:243–264.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Baegin Seong. 1977. Romanization of the special letters of manchu. *Eoneohag*, 2:185–197.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Reddy Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. *CoRR*, abs/2005.04816.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Tasaku Tsunoda. 2006. *Language endangerment and language revitalization: An introduction*. De Gruyter Mouton.

Alexander Vovin. 2023. Written Manchu. In Alexander Vovin, José Andrés Alonso de la Fuente, and Juha Janhunen, editors, *The Tungusic Languages*, pages 103–138. Routledge, New York.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Hyun-Jo You. 2014. A manchu speller: With a practical introduction to the natural language processing of minority languages. *Altai Hakpo*, 24:39–67.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. Chren: Cherokee-english machine translation for endangered language revitalization.

A Example Appendix

<Manchu sentence>

ᠴᠣᠣᡥᠠ ᠪᡝ ᠸᠠᡴᡳ ᠰᡝᠮᡝ ᡨᡠᠮᡝᠨ ᠴᠣᠣᡥᠠ ᠪᡝ ᡠᠩᡤᡳᡶᡳ ᡨᠣᠰᠣᡥᠣ᠈

cooha be waki seme tumen cooha be unggifi tosoho,

<Translated sentence>

군사를 죽이려고 군사 일만 명을 보내서 길을 막았다.

kwunsalul cwukilyeko kwunsa ilman myengul ponayse kilul makassta

Figure 2: Example of Romanizations of Manchu text and Korean text