Bridging the Linguistic Divide: Developing a North-South Korean Parallel Corpus for Machine Translation

Hannah Hyesun Chun[1], Chanju Lee[1], Hyunkyoo Choi[2], Charmgil Hong[1]
[1]Handong Global University
[2]Korea Institute of Science and Technology Information
{22000662, 21800587, charmgil}@handong.ac.kr, hkchoi@kisti.re.kr

Abstract

This study addresses a significant challenge in machine translation between North and South Korean languages: the scarcity of parallel corpora. To overcome this limitation, we developed a comprehensive North-South Korean parallel corpus and fine-tuned a South Korean pre-trained model. Our research explores the potential for a robust sentence-level translation model between the two Korean dialects. We evaluated the performance of the model using both BLEU and BERTScore metrics and conducted a qualitative analysis to assess its ability to capture the distinct linguistic features of North and South Korean languages, including differences in vocabulary, word spacing, and spelling. Our findings demonstrate that this newly developed corpus and translation model not only enhance machine translation capabilities but also contribute valuable insights to linguistic studies of the two Korean languages.

1 Introduction

Korean is the official language of both South Korea and North Korea. Because of the geographical and sociopolitical division of the Korean peninsula for over 70 years the Korean language has evolved differently in the two Koreas. The most notable difference can be found in the vocabulary: everyday North Korean terms differ by 38% from those used in South Korea, while technical terms differ by 66% (Park, 2016). Additionally, differences in orthography and discourse style often prevent North and South Korean speakers from understanding each other. According to the *2016 Survey on Language Awareness of North and South Korea* by the *National Institute of the Korean Language*, 29.8% of North Korean defectors need 4 to 5 years to speak and write like South Koreans, while 51% need more than 6 years (National Institute of Korean Language, 2016). This language gap between North and South Korea could pose a practical obstacle to Korean reunification.

Efforts have been made to overcome the linguistic divide between North and South Korea. For instance, in 2005, both countries undertook a joint project, to create a unified Korean dictionary, Gyeoremal-keunsajeon (Yu, 2021). Unfortunately, the project was discontinued due to turbulent inter-Korean relations, with only about 40% of the total 307,000 words collaboratively discussed and resolved (Park, 2023). Additionally, in South Korea, a translator app, *Geul-dong-mu* was launched to help North Korean defectors adjust to their new lives by translating South Korean terms into North Korean equivalents (Geuldongmu). However, the app had a limited vocabulary and could not translate sentences, which highlights the need for more natural language processing (NLP) research focused on the North Korean language. The paucity of North Korean language resources has made it challenging to build a large-scale corpus for NLP tasks like machine translation in the North Korean language.

Several studies have implemented NLP research on the North Korean language. For example, (Kim et al., 2022) created North Korean-English and North Korean-Japanese parallel corpora from a North Korean News portal, *Uriminzokkiri*. This parallel corpus was used to conduct North Korean translation experiments. Another study (Akdemir et al., 2022) built a North Korean corpus using *Rodong News articles* and *New Year Addresses of the North Korean leaders* to train a BERT-based language model and a sentiment analyzer. However, these studies were limited to corpora that either only included North Korean language data or were paired with languages like Japanese or English. Studies focusing on translation between North and South Korean using a parallel corpus that align North Korean sentences with their South Korean counterparts are scarce.

To address this problem, we created a parallel

corpus by collecting and aligning text data, which are translations between North Korean and South Korean. We then developed a North-South Korean machine translation model by fine-tuning a South Korean pre-trained model with the parallel corpus. We conducted a quantitative evaluation using combined metrics: BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). Furthermore, we performed a thorough qualitative analysis of our translation results to assess the extent to which our translation model captures the differences between the North Korean and South Korean languages.

2 Related Works

NLP research on the North Korean language is limited due to the scarcity of North Korean NLP tools and resources. Consequently, North Korean-related NLP research lags behind in reaching cutting-edge results.

One study (Kim et al., 2023) constructed a parallel corpus specifically for the North Korean neural machine translation (NMT) systems. Using a news portal called *Uriminzokkiri*, news articles published in North Korean, English, and Japanese were aligned manually and automatically to create North Korean–English (NK-EN) and North Korean–Japanese (NK-JA) parallel corpora. A trilingual annotator manually aligned the English and Japanese sentences with their corresponding North Korean sentences to create evaluation data for machine translation. The automatic alignment method was used to generate the training data, and the study found that the bidirectional South Korean NMT model (bidi-SK) achieved the highest alignment score. Translation experiments using the NK-EN and NK-JA datasets showed that the South Korean pre-trained BART fine-tuned with North Korean data produced a higher BLEU score than the models trained exclusively on either South or North Korean data. Similar to the study, we utilize a South Korean pre-trained BART-based model. However, unlike that study, we focus on exploring the potential for machine translation between North and South Korean languages.

Another study (Choi and Hong, 2022) constructed a South Korean-North Korean parallel corpus to develop a North and South Korean bidirectional translator. South Korean-English and North Korean-English news data from the Korean Parallel Data (Park et al., 2016) were used as resources. Corresponding South Korean and North Korean sentences for each English sentence were extracted and aligned to create a parallel dataset. However, due to the small size of the training data, only a few linguistic differences were observed in the dataset and the translation results.

In our study, we created a large and diverse dataset of North Korean and South Korean languages by incorporating novels with varied content, sentence structures, and vocabulary. We then fine-tuned the South Korean BART model [1] using this dataset to improve the translation quality between North Korean and South Korean languages.

3 North – South Korean Parallel Corpus Construction

3.1 Data Description

Table 01 describes the information about the North-South Korean parallel corpus. The corpus contains 130,738 sentence pairs, sourced from classic novels and the Bible, translated into North and South Korean languages. 29,986 sentence pairs were created from the Bible, making up 23% of the corpus. Meanwhile, classic novels provided 100,752 sentence pairs, representing 77% of the corpus. English and French classic novels comprise 72% of the dataset, while Korean classic novels account for 5%. The parallel corpus is available in GitHub to encourage further research in Natural Language Processing(NLP) or Linguistics. [2]

3.2 Data Collection

The dataset was created using the North Korean and South Korean versions of the Bible, three Korean classic novels, and one English and French classic novel, each translated into North Korean and South Korean. The Bible was selected because of its consistent structure of books, chapters, and verses, which made it easier to match corresponding sentences across translations. The North Korean version of the Bible was kindly provided by the *North Korean Science and Technology Network (NKTech)* of the *Korea Institute of Science and Technology Information*. The Korean, English, and French classic novels were chosen for their variety of everyday words, discourse styles, and sentence structures. The North Korean versions of the classic novels were acquired from the *Information Center on North Korea*, operated by the *Ministry of*

[1] https://github.com/SKT-AI/KoBART

[2] https://github.com/HandongSF/KoreanUnificationParallelCorpus

Resource	Title	Sentence pairs	Total sentence pairs
Bible		29,986	29,986
English Classic Novel	Jane Eyre	60,331	94,459
French Classic Novel	The Red and the Black	34,128	
Korean Classic Novel	Onggojip-jeon (옹고집전)	988	6,293
	Sukhyang-jeon (숙향전)	3,538	
	Shimchung-jeon(심청전)	1,767	
			130,738

Table 1: North-South Korean parallel corpus

Unification in South Korea. All documents, including the matching South Korean translations, were manually collected in PDF format using scanning software. They were then converted into text using an optical character recognition (OCR) tool and carefully reviewed to correct any typos or errors during the automated recognition process.

3.3 Manual error correction of the text data

The accuracy of the extracted text data was cross-checked with the original PDF, and any spelling or spacing mistakes were corrected. Annotations, chapter titles, page numbers, Chinese characters, and languages other than North or South Korean were eliminated. A unique challenge was to avoid unintentional modification of North Korean text according to South Korean language rules.

3.4 Matching sentence pairs

The resulting text files were preprocessed to remove all punctuation marks except periods (.), commas (,), question marks (?), and exclamation marks (!). These four punctuation marks were used to separate the text into sentences. The text files were then converted into a spreadsheet with North Korean sentences in the first column ('nk') and South Korean sentences in the second column ('sk').

Next, matching was performed to pair each North Korean and South Korean sentence with an identical meaning. In the case of classic novels, one North Korean sentence often corresponded to multiple South Korean sentences, and *vice versa*. Multiple sentences were allowed in the same row to create sentence pairs as long as one or more sentences in both languages shared the same meaning. Any sentences with no corresponding match in the other language were deleted.

These stages of building the parallel corpus required significant labor. Approximately twenty participants were recruited to handle routine tasks, such as applying the OCR tool and identifying obvious errors and mistakes. All participants were native South Korean speakers, with no specific criteria regarding age, gender, major, or grade in the selection process. Each part of the dataset was reviewed twice by different participants to minimize individual bias and ensure consistency when pairing North Korean and South Korean sentences with equivalent meanings. The initial completion of the dataset took approximately six weeks, from August 29th to October 6th, 2023.

In the resulting dataset, errors were more prevalent in the North Korean text than in the South Korean text, as the process was conducted by native South Koreans with little or no knowledge of the North Korean language. Therefore, an additional phase was undertaken to correct the errors in the North Korean text. Important spelling, spacing, and vocabulary differences between the North and South Korean languages were studied in advance to reduce the likelihood of overlooking errors. In total, the creation of the North-South Korean parallel corpus took about three to four months.

4 North–South Korean Translation Experiments

Using the North-South Korean parallel corpus constructed in Section 3, we trained a North-South Korean bidirectional translation model and measured the translation quality with BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). To evaluate how well the model captured the linguistic differences between North and South Korean, we analyzed the translation results of several sentences that were not included in the training data.

4.1 Experimental Setup

Dataset

For the translation experiments, we split the North-South Korean parallel corpus, consisting of

130,738 sentence pairs, into training and test sets with a 9:1 ratio. To ensure balanced representation, the test set of 13,073 sentence pairs was proportionally collected as follows: The Bible 23% (3,007 pairs), *Jane Eyre* 46% (6,016 pairs), *The Red and the Black* 26% (3,399 pairs), and Korean classic novels 5% (651 pairs), with the sentence pairs selected randomly. The remaining sentences were used for the training set. For *Jane Eyre* and *The Red and the Black*, the number of South Korean publishers was also considered. Since the South Korean translations of these works were provided by multiple publishers, some sentence pairs had the same North Korean sentence aligned with different versions of South Korean sentences. Therefore, extra caution was needed to make sure that North Korean sentences in the training set were not included in the test set. For example, because *Jane Eyre* was sourced from four South Korean publishers, around 1,500 North Korean sentences were selected from each publisher to make up the 6,016 sentence pairs needed for the test set. A similar approach was applied to *The Red and the Black*.

Baseline Model

BART (Lewis et al., 2019) is a denoising autoencoder that learns to map corrupted sentences to their original forms and has achieved high performance in various text generation tasks. It has shown enhanced performance in Romanian-English translation. KoBART is a South Korean BART trained on approximately 40GB of Korean text. Fine-tuning KoBART has proven effective in improving North Korean-Japanese and North Korean-English machine translation (Kim et al., 2023). Therefore, we chose KoBART as our baseline model. Specifically, the KoBART-translation model was fine-tuned on our dataset using the following hyperparameters: a batch size of 4, 8 epochs, a learning rate of 3e-5, and the AdamW optimizer. Sentences were pre-processed using the KoBART tokenizer.[3] We refer to the model translating North Korean sentences to South Korean as NK→SK, and South Korean sentences to North Korean as SK→NK.

4.2 Experimental Results

The translation performance of the NK→SK and SK→NK models was evaluated using two metrics: BLEU Score (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). Table 02 presents the BLEU scores for each model. The NK→SK model achieved a score of 0.442, outperforming the SK→NK model, which scored 0.107. We attribute the higher score of the NK→SK model to the difference in the number of reference sentences possible for comparison with the output of the model. That is, since the South Korean sentences were drawn from various publishers, the NK→SK model had at most four reference sentences to compare against each translation output. In contrast, the SK→NK model had only one reference North Korean sentence available for each translation output, as the North Korean sentences were extracted from a single publisher.

Model	BLEU Score	BERT Score
NK→SK	0.442	0.821
SK→NK	0.107	0.815

Table 2: BLEU Score and BERT Score of the NK→SK and the SK→NK model

Although BLEU is a commonly used evaluation metric in machine translation tasks, it relies on surface-form similarity measures and often neglects semantic equivalence between the reference and candidate. This can lead to underestimating the performance of semantically correct translations that differ from the surface form of the reference. To address the limitations of BLEU, we used BERTScore as an additional metric. Unlike n-gram-based metrics, BERTScore measures the cosine similarity between tokens in the candidate and reference using the contextual embeddings of BERT. BERTScore correlates better with human judgment as it effectively captures the semantic and contextual information of words and phrases in the candidate and reference. BERTScore calculates precision, recall, and F1 scores. Recall measures how well each token in the reference is captured by those in the candidate, while precision measures how closely candidate tokens match those in the reference. F1 is the harmonic mean of precision and recall. Table 02 presents the BERTScores for the NK→SK and SK→NK models, with only the F1 scores recorded for simplicity. Both models reached high scores, with the NK→SK model scoring 0.821 and the SK→NK model scoring 0.815. The difference in the BERTScore results between the two models is much smaller than the difference observed in their BLEU scores, indicating

[3] https://github.com/SKT-AI/KoBART?tab=readme-ov-file#tokenizer

Source	탁자우의 나의 명함이 나의 이름을 **말밥에 올리**는것이였소.
NK→SK	테이블 위의 내 명함이 내 이름을 **화제에 올려**놓았소.
English	A card of mine lay on the table; this being perceived, **brought my name under discussion**.
Source	레날부인은 이렇게 운명의 희롱으로 훌렁 빠져들어간 이 무서운 정열의 괴롬으로 **모대기**고있었다.
NK→SK	레날 부인은 운명의 희롱에 걸려든 이 끔찍한 정열의 고통에 **사로잡혀** 있었다.
English	Madame de Rênal was **a prey to all the poignancy** of the terrible passion in which chance had involved her.

Table 3: Example of NK→SK translations that show North and South Korean difference in vocabulary usage

Source	그래 열린 창문으로 손을 디밀어 **창가림**을 치고 안을 들여다볼수 있을만큼 틈새를 남겨놓았소.
NK→SK	열린 창문으로 손을 집어넣어 **커튼**을 젖히고 안을 들여다볼 수 있을 만큼만 틈을 남겨놓았소.
Source	그러고는 열려 있는 창문 틈으로 손을 넣어서 창문 위로 **커튼**을 치고 안을 살펴볼 수 있을 만큼만 공간을 남겨 두었소.
SK→NK	그리고는 열려있는 창문틈으로 손을 집어넣어 창문에 **창가림**을 치고 안을 들여다보게 하였다.
English	So putting my hand in through the open window, I drew the **curtain** over it, leaving only an opening through which I could take observations.

Table 4: Example of NK→SK and SK→NK translations that show North and South Korean difference in loanwords

no significant performance difference between the NK→SK and SK→NK models regarding semantic similarity.

4.3 Qualitative Analysis

A qualitative analysis was conducted to evaluate the ability of the translation model, considering the differences between the North and South Korean languages. The analysis focused on three main aspects: vocabulary, spelling, and word spacing. These criteria for comparing North and South Korean language differences were chosen based on the book, *Understanding the Languages of North and South Korea* (Cho et al., 2002). Sentence pairs that contained the key linguistic differences between North and South Korean languages were selected from the test dataset.

4.3.1 Vocabulary difference between North and South Korea

North Korea and South Korea often use different words to refer to the same meaning. From a native South Korean speaker's perspective, some North Korean words are difficult to understand without knowing the North Korean language.

For instance, the North Korean phrase "말밥에 오르다(mal-bab-e o-leu-da)" means "being spoken about by many people." Not only "말밥(mal-bab)" is an unfamiliar word in South Korea, but rarely used with the predicate "오르다(o-leu-da)," which means "to come up." Instead, a more commonly used phrase in South Korean is "화제에 오르다(hwa-je-e o-leu-da)." For this reason, we selected a North Korean sentence including the phrase "말밥에 오르다(mal-bab-e o-leu-da)" to check the translation results of the NK→SK model. Table03 shows that the NK→SK model successfully translated the North Korean phrase "말밥에 오르다(mal-bab-e o-leu-da)" into the more generally used South Korean phrase "화제에 오르다(hwa-je-e o-leu-da)."

Another example is the North Korean verb "모대기다(mo-dae-gi-da)" which means "to writhe in agony." In South Korean, "모대기다(mo-dae-gi-da)" does not convey the same meaning and is unused. So, we chose a North Korean sentence with the phrase "모대기다(mo-dae-gi-da)" to assess the translation results of the NK→SK model. In Table03, the NK→SK model translated the North Korean phrase "모대기다(mo-dae-gi-da)" into a South Korean phrase "사로잡히다(sa-lo-chap-hi-ta)," which means "to be dominated or

Source	요셉은 바로의 명령대로 그들에게 마차를 내여주고 **려행**길에 먹을 량식도 마련해주었다.
NK→SK	요셉은 파라오가 시키는 대로 했습니다. 요셉은 짐마차와 **여행**길에 필요한 음식을 그들에게 주었습니다.
Source	요셉은 파라오가 명령한 대로 그들에게 수레 몇 대를 주었고 **여행**할 때 필요한 음식도 주었습니다.
SK→NK	요셉은 바로가 지시한대로 수레를 몇대 주고 **려행**할 때 먹을것을 주었다.
English	Joseph gave them carts, as Pharaoh had commanded, and he also gave them provisions for their **journey**.

Table 5: Example of NK→SK and SK→NK translations that show North and South Korean difference in the initial sound rule

Source	그 빛은 또한 **난로가**에 모여앉은 한패의 사람들도 밝히고있었다.
NK→SK	그 빛은 또 **난롯가**에 모여 있는 한 무리의 사람들을 밝혀 주었다.
Source	그 빛은 **난롯가** 곁에 모여 앉은 사람들을 아늑하게 비춰 주고 있었다.
SK→NK	그 빛은 **난로가**에 앉아있는 사람들을 환하게 비쳐주었다.
English	This ruddy shine revealed, too, a group near the **mantelpiece**.

Table 6: Example of NK→SK and SK→NK translations that show North and South Korean difference in the addition of a "ㅅ" into a compound word

overwhelmed by a particular emotion." Although it is not a perfect match, the meaning of the North Korean phrase is communicated to a certain degree.

Loanwords

English loanwords are one of the noticeable differences in vocabulary between North and South Korean languages. North Korea uses less foreign loanwords than South Korea. One example is the English word "curtain," which is rendered as the loanword "커튼(keo-teun)" in South Korea, whereas in North Korea, it is referred to as "창가림(chang-ga-lim)." "창가림(chang-ga-lim)" joins together the Korean words "창(chang)," meaning "window," and "가림(ga-lim)," meaning "cover." Table04 shows that the NK→SK model accurately translated the North Korean word "창가림(chang-ga-lim)" into the South Korean word "커튼(keo-teun)." It also shows the example of the SK→NK model correctly translating "커튼(keo-teun)" into "창가림(chang-ga-lim)."

4.3.2 Spelling difference between North and South Korean

Initial sound rule

In South Korean, the consonant "ㄹ" changes into "ㅇ" or "ㄴ" when combined with vowels "ㅑ, ㅕ, ㅛ, ㅠ, ㅢ, ㅣ, ㅖ," while North Korean does not undergo such changes. For example, the word "travel" is written as "려행(lyeo-haeng)" in North Korean but "여행(yeo-haeng)" in South Korean. In Table 05, we see that the NK→SK model correctly translated the North Korean word "려행(lyeo-haeng)" to "여행(yeo-haeng)" in South Korean, while the SK→NK model translated "여행(yeo-haeng)" into "려행(lyeo-haeng)" in North Korean, according to the initial sound rule.

The addition of "ㅅ" into a compound word

In South Korean, the consonant "ㅅ" is added when compound words are created in certain cases. The first case involves compound words made up of pure Korean words, such as "시냇가(si-naet-ga)," meaning "the surroundings of a stream." Since both "시내(si-nae)," meaning "stream," and "가(ga)," meaning "the surroundings," are two pure Korean words with no corresponding Chinese characters, "ㅅ" should be added when those words are combined to form a new word. Another case occurs when compound words consist of one pure Korean word and another Sino-Korean word, with the first word ending in a vowel. An example is "난롯가(nal-lok-ga)," which means "fireside." "난로(暖爐)" is a Sino-Korean word meaning "stove" and ends in a vowel, "ㅗ," requiring a

Source	그래 나는 상당한 돈을 주고 적당한 일자리를 얻어준 다음에 체면을 유지하며 기꺼이 **헤여졌소**.
NK→SK	그래서 상당한 돈을 주고 적당한 일자리를 얻어 준 다음부터 나는 체면을 되찾고 기꺼이 **헤어졌소**.
Source	결국 충분한 돈을 주어 장사를 시작하게 해주고, 깨끗이 **헤어지고** 나니 마음이 후련했소.
SK→NK	마침내 저에게 충분한 돈을 지불하고 그와 **헤여져** 있게 되자 나는 무척 안도감을 느꼈소.
English	I was glad to give her a sufficient sum to set her up in a good line of business, and so **get decently rid of her**.

Table 7: Example of NK→SK and SK→NK translations that show North and South Korean difference in the endings of a word "-아/-어" based on the final syllable vowel of the stem

Source	그 옷은 세상의 어떤 빨래하는 사람도 그보다 더 희게 할수 **없을만큼** 새하얗고 눈부시게 빛났다.
NK→SK	그 옷은 세상 어느 누구도 그보다 더 희게 할 수 **없을 만큼** 새하얗고 눈부시게 빛났다.
English	His clothes became dazzling white, **whiter than anyone** in the world could bleach them.
Source	하느님 당신의 길은 거룩하시오니 **하느님만큼** 높은 신이 어디 있으리이까.
NK→SK	오 하나님 주의 길은 거룩합니다. **하나님만큼** 위대한 신이 어디 있습니까?
English	Your ways, God, are holy. What god is **as great as our God**?

Table 8: Example of NK→SK translations that show North and South Korean difference distinguishing word spacing rules between dependent nouns and particles

consonant "ㅅ" to be added when combined with the word "가(ga)." North Korean, on the other hand, does not apply this rule. Hence, the South Korean word "시냇가(si-naet-ga)" and "난롯가(nal-lok-ga)" are written as "시내가(si-nae-ga)," and "난로가(nal-lo-ga)," respectively. Table 06 demonstrates an example of the NK→SK model correctly translating "난로가(nal-lo-ga)" into the South Korean word "난롯가(nal-lok-ga)," following the rule of adding "ㅅ" in compound words. The SK→NK model correctly translated "난롯가(nal-lok-ga)" into the correct North Korean spelling, "난로가(nal-lo-ga)."

Endings of a word "-아/-어" based on the final syllable vowel of the stem

South Korean and North Korean differ in the way of writing the ending of a word "-아/-어" depending on the final syllable vowel of the stem. In North Korean, the ending of a word is written as "-여/-였" when the final syllable vowel of the stem is "ㅣ, ㅐ, ㅖ, ㅚ, ㅟ, ㅢ" and "하". In South Korean, the ending of the word is written as "-아" when the final syllable vowel of the stem is "ㅏ, ㅗ," and "-어" otherwise. For example, "to break off a relationship" is written "헤어지다(he-eo-ji-da)" in South Korean because the final syllable vowel of the stem "헤" is "ㅖ," not either "ㅏ" or "ㅗ." In North Korean, since "ㅖ" is in one of the vowels listed(ㅣ, ㅐ, ㅖ, ㅚ, ㅟ, ㅢ), they write "헤여지다(he-yeo-ji-da)." Table 07 illustrates an example where the NK→SK model precisely translated "헤여지다(he-yeo-ji-da)" into the South Korean word "헤어지다(he-eo-ji-da)." The SK→NK model correctly translated "헤어지다(he-eo-ji-da)" into the North Korean word "헤여지다(he-yeo-ji-da)".

4.3.3 Word spacing difference between North and South Korean

Spacing of dependent nouns and particles

North and South Korean have different word spacing rules for dependent nouns and particles. In South Korean, the dependent noun is separated from the preceding verb or adjective stem, while in North Korean, it is attached. However, both languages attach particles to the preceding word. Therefore, we checked whether the NK→SK model could distinguish between dependent nouns and a particle that looks identical and apply the correct word spacing rules accordingly. Table 08 provides an example of the NK→SK model successfully translating "만큼(man-keum)," when used as both a dependent noun and a particle. When "만큼(man-keum)" is used after the

Source	...네 시어미 무슨 일로 통곡하는지 아구리를 당장 다물지 않으면 **쫓아낸다**고 일러라.
SK→NK	...네 시어미 무슨 일로 통곡하는지 아구리를 당장 다물지 않으면 **쫓아낸다**고 일러라.
Source	...네 시어미 무슨 일로 통곡하는지 아구리를 당장 다물지 않으면 **쫓아 낸다**고 일러라.
SK→NK	...네 시어미 무슨 일로 통곡하는지 아구리를 당장 다물지 않으면 **쫓아낸다**고 일러라.
English	Tell your mother-in-law that if she doesn't stop wailing this instance, I'll **throw her out**.

Table 9: Example of SK→NK translations that show North and South Korean difference applying spacing rules between main and auxiliary predicate element.

stem of the adjective "없을(eobs-eul)," meaning "something does not exist," it is a dependent noun and translated into South Korean with a word space between the two words. On the contrary, when "만큼(man-keum)" comes after the noun "하느님(ha-neu-nim)," meaning "God," it is a particle and attached to the previous word when translated into South Korean.

Spacing between predicate elements
North and South Korean also differ in the rules for spacing between main and auxiliary predicate elements. In North Korean, the main and the auxiliary predicate elements are always attached. In South Korean, separating the main and auxiliary predicates is a general rule. However, it is also possible to attach the auxiliary predicates to the main predicate in some cases. One such case is when a main predicate whose final syllable vowel is "아, 어, 여" is followed by certain auxiliary predicates, such as "내다(nae-da)." "쫓아내다(jjoch-a-nae-da)," which means to "drive somebody out," is an example. In contrast to North Korean, South Korean allows both "쫓아내다" and "쫓아 내다" because "쫓아" ends with "아" and precedes the predicate "내다". Table 09 shows a correct translation example of the SK→NK model translating both "쫓아 내다" and "쫓아내다" into "쫓아내다" in North Korean.

5 Conclusion

This study presented a North-South Korean parallel corpus using the Bible and literature resources. This corpus is particularly significant because a sizable parallel corpora containing North and South Korean sentence pairs is scarce. Additionally, the Bible and literary texts offer a diverse range of content and sentence structures. The resulting corpus was then used to train and analyze a North-South Korean bidirectional translation model. The translation quality of the model was quantitatively evaluated using two metrics: BERTScore and BLEU. The BERTScore results show that the NK→SK and SK→NK models achieved high translation performance on the test set. We conducted an in-depth qualitative analysis of the translation results, focusing on linguistic differences between the North and South Korean languages in three key areas: vocabulary, spelling, and spacing. Our findings demonstrated that fine-tuning a South Korean pre-trained model with the North-South Korean parallel corpus can produce a translation model capable of accurately translating sentences between the two languages. One drawback of our parallel corpus is the lack of sentences including contemporary loanwords and technical terms, due to the historical period of the literary resources and the Bible. For this reason, our qualitative analysis of the translation model has limitations in assessing the translation quality of sentences with recent loanwords and terms primarily used in a professional field. Therefore, expanding the North-South parallel corpus with sentence pairs from various sources, such as research papers, late 20th or 21st-century literature, or movie subtitles, is necessary. Moreover, we plan to investigate methods and large language models that can further improve the performance of the machine translation between the two languages.

Acknowledgments

This research was supported (1) by the Korea Institute of Science and Technology Information (K-23-L01-C01, Construction on Intelligent SciTech Information Curation), (2) by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (RS-2024-00431394) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation), and (3) by the MSIT, Korea, under the National Program for Excellence in SW, supervised by the IITP (2023-0-00055).

References

Arda Akdemir, Yeojoo Jeon, and Tetsuo Shibuya. 2022. Developing language resources and nlp tools for the north korean language. *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 5595–5600.

Ohyeon Cho, Yonggyeong Kim, and Donggeun Park. 2002. *남북한 언어의 이해 Understanding the Languages of North and South Korea*. Youkrack.

Hoyoon Choi and Charmgil Hong. 2022. Neural machine translation using south korean-north korean parallel corpus. *Proceedings of Korea Multimedia Society Conference.*, 25.

Geuldongmu. 2015. Geuldongmu: North-south korean translator. `https://www.geuldongmu.org/`. Official website of Geuldongmu.

Hwichan Kim, Sangwhan Moon, Naoaki Okazaki, and Mamoru Komachi. 2022. Learning how to translate north korean through south korean. *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 6711–6718.

Hwichan Kim, Hirasawa Tosho, Sangwhan Moon, Naoaki Okazaki, and Mamoru Komachi. 2023. North korean neural machine translation through south korean resources. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871–7880.

National Institute of Korean Language. 2016. 2016 survey on language awareness of north and south korea. *National Institute of the Korean Language*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Byeong-Yong Park. 2016. 남북한 언어 차이 심각… 일반어 38%, 전문어 66% 달라 Serious language difference between North and South Korea… 38% difference for everyday language, 66% for specialized language. *VoA Korea*.

Ga-Young Park. 2023. Waiting on the north: Unified korean dictionary project's long journey. *The Korea Herald*.

Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. 2016. Korean language resources for everyone. *In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*.

Hyun-Kyung Yu. 2021. Gyeoremal-keunsajeon as a dictionary of language integration in south and north korea. *Yonsei University Institute of Language and Information Studies*, 53:5–30.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *8th International Conference on Learning Representations*.