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

This research was supported by
(1) the Korea Institute of Science and Technology Information (KISTI) in 2023 (K-23-L01-C01, Construction on Intelligent SciTech Information Curation)
(2) 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)
(3) the MSIT, Korea, under the National Program for Excellence in SW, supervised by the IITP in 2024 (2023-0-00055)

Table of Contents

1. Introduction
2. Related Works
3. Contributions
4. North – South Korean Parallel Corpus Construction
5. North – South Korean Translation Model Training and Evaluation
6. Conclusion
7. Future Work

Introduction

[4]

- **Division of Korea □ Differences between North and South Korean Languages**
  - **Everyday** North Korean **terms** differ by **38%** from those used in South Korea, while **technical terms** differ by **66%**[1]
  - **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**[2]
- **Studies** focusing on **machine translation between North and South Korean are rare due to the scarcity of a North-South Korean parallel corpus**[3]

[5]

	Shampoo	Friend
	샴푸 (Shampu)	친구 (Chingu)
	머리물비누 (Meorimulbinu)	동무 (Dongmu)

[1] "Serious language difference between North and South Korea 38% difference for everyday language, 66% for specialized language", SBS news, 2016.
[2] "2016 Survey on Language Awareness of North and South Korea", 국립국어원, 2016.
[3] Hwichan Kim et al., "Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
[4] "South&North Korea map"
[5] "Crossing Divides: Two Koreas divided by a fractured language", BBC news, 2019.

Related Works

1. ***North Korean Neural Machine Translation through South Korean Resources*** [1]
   - **North Korean→English**, **North Korean→Japanese machine translation** experiments conducted
   - From **a North Korean news portal**, *Uriminzokkiri*, **news articles** published in **North Korean** and translated into **English** and **Japanese** were aligned to create a total of **5,000 North Korean–English** sentence pairs and a total of **4,700 North Korean–Japanese** sentence pairs

[1] Hwichan Kim et al., "North Korean Neural Machine Translation through South Korean Resources", ACM Transactions on Asian and Low-Resource Language Information Processing, 2023.

Related Works

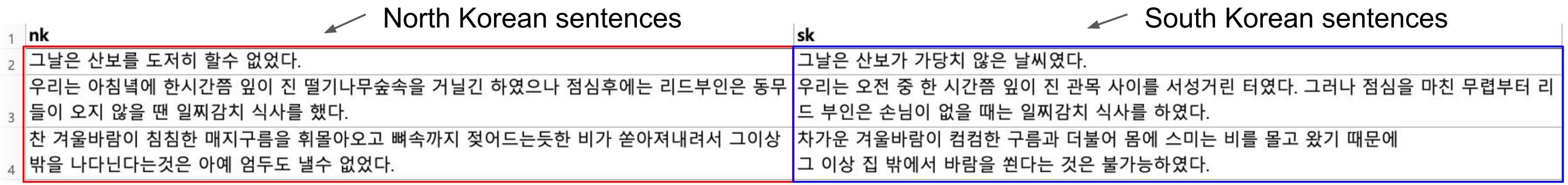
2. *Neural machine translation using south korean-north korean parallel corpus* [1]

- From **Korean Parallel Data,**[2] **"South Korean-English", "North Korean-English" news data** were aligned to create a total of 3,000 **"South Korean-North Korean" sentence pairs**
- Due to the **small size** of the **training data**, only **a few linguistic differences** were observed in the **dataset** and the translation results

[1] Hoyoon Choi et al., "Neural Machine Translation using South Korean-North Korean Parallel Corpus", *Proceedings of Korea Multimedia Society Conference*, 2022.
[2] Jungyeul Park et al., "Korean Language Resources for Everyone", In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, 2016.

Contributions

- **Construct a North-South Korean parallel corpus using the North and South Korean versions of the Bible, Classic Novels, and Korean Folk Tales**
- Release the North-South Korean parallel corpus for further research
- Prove the efficacy of the North-South Korean parallel corpus by developing and evaluating a North-South Korean bidirectional translation model

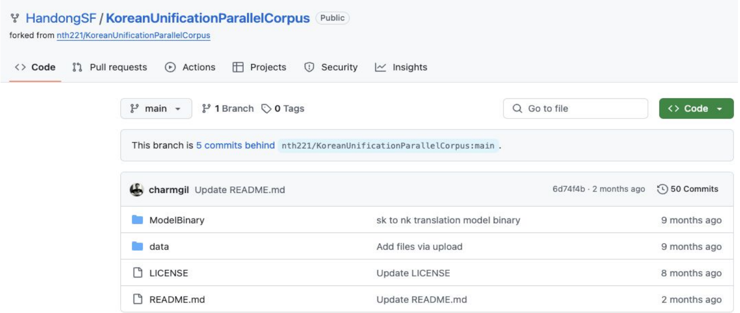


1	nk	sk
2	그날은 산보를 도저히 할수 없었다.	그날은 산보가 가당치 않은 날씨였다.
3	우리는 아침녘에 한시간쯤 잎이 진 떨기나무숲속을 거닐긴 하였으나 점심후에는 리드부인은 동무들이 오지 않을 땐 일찌감치 식사를 했다.	우리는 오전 중 한 시간쯤 잎이 진 관목 사이를 서성거린 터였다. 그러나 점심을 마친 무렵부터 리드 부인은 손님이 없을 때는 일찌감치 식사를 하였다.
4	찬 겨울바람이 침침한 매지구름을 휘몰아오고 뼈속까지 젖어드는듯한 비가 쏟아져내려서 그이상 밖을 나다닌다는것은 아예 엄두도 낼수 없었다.	차가운 겨울바람이 컴컴한 구름과 더불어 몸에 스미는 비를 몰고 왔기 때문에 그 이상 집 밖에서 바람을 쐰다는 것은 불가능하였다.

an example of the sentence pairs in the North-South Korean Parallel Corpus

Contributions

- Construct a North-South Korean parallel corpus using the North and South Korean versions of the Bible, Classic Novels, and Korean Folk Tales
- **Release the North-South Korean parallel corpus for further research**
- Prove the efficacy of the North-South Korean parallel corpus by developing and evaluating a North-South Korean bidirectional translation model



HandongSF / KoreanUnificationParallelCorpus

forked from nth221/KoreanUnificationParallelCorpus

This branch is 5 commits behind `nth221/KoreanUnificationParallelCorpus:main` .

charmgil Update README.md 6d74f4b · 2 months ago 50 Commits

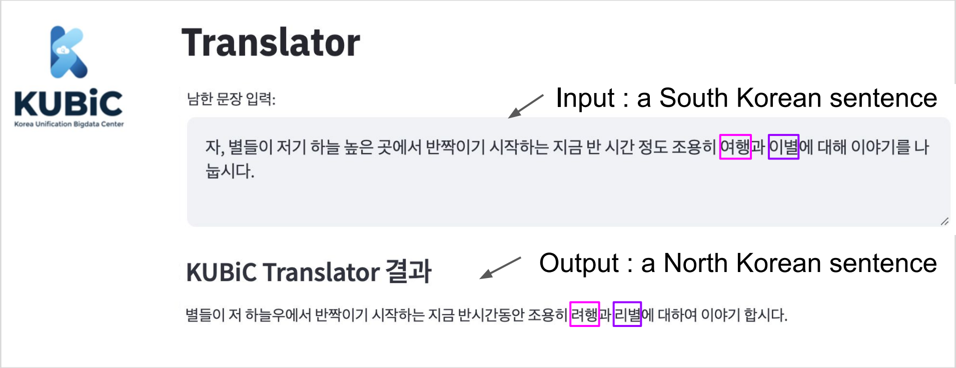
ModelBinary	sk to nk translation model binary	9 months ago
data	Add files via upload	9 months ago
LICENSE	Update LICENSE	8 months ago
README.md	Update README.md	2 months ago

the North-South Korean Parallel Corpus published to Github

https://github.com/HandongSF/KoreanUnificationParallelCorpus

Contributions

- Construct a North-South Korean parallel corpus using the North and South Korean versions of the Bible, Classic Novels, and Korean Folk Tales
- Release the North-South Korean parallel corpus for further research
- **Prove the efficacy of the North-South Korean parallel corpus by developing and evaluating a North-South Korean bidirectional translation model**



Translator

남한 문장 입력:

Input : a South Korean sentence

자, 별들이 저기 하늘 높은 곳에서 반짝이기 시작하는 지금 반 시간 정도 조용히 여행과 이별에 대해 이야기를 나눕시다.

KUBiC Translator 결과

Output : a North Korean sentence

별들이 저 하늘우에서 반짝이기 시작하는 지금 반시간동안 조용히 려행과 리별에 대하여 이야기 합시다.

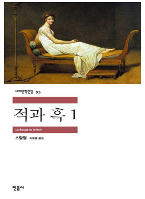
an example translation output of the SK → NK model

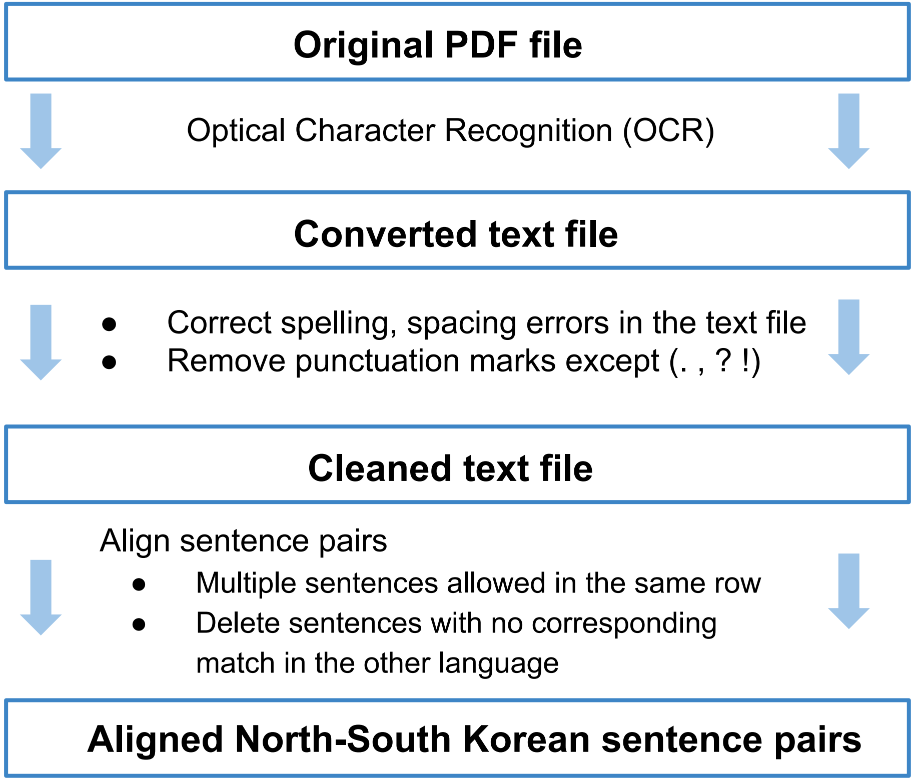
North-South Korean Parallel Corpus

North – South Korean Parallel Corpus Construction

Information Center on North Korea operated by the *Ministry of Unification* in South Korea

- English Classic Novel
- French Classic Novel
- Korean Folk Tales





Original PDF file

Optical Character Recognition (OCR)

Converted text file

- Correct spelling, spacing errors in the text file
- Remove punctuation marks except (. , ? !)

Cleaned text file

Align sentence pairs

- Multiple sentences allowed in the same row
- Delete sentences with no corresponding match in the other language

Aligned North-South Korean sentence pairs

North Korean Science and Technology Network of the *Korea Institute of Science and Technology Information*

- Bible

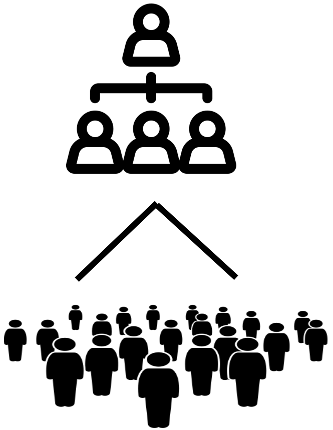
구약전서

신약전서

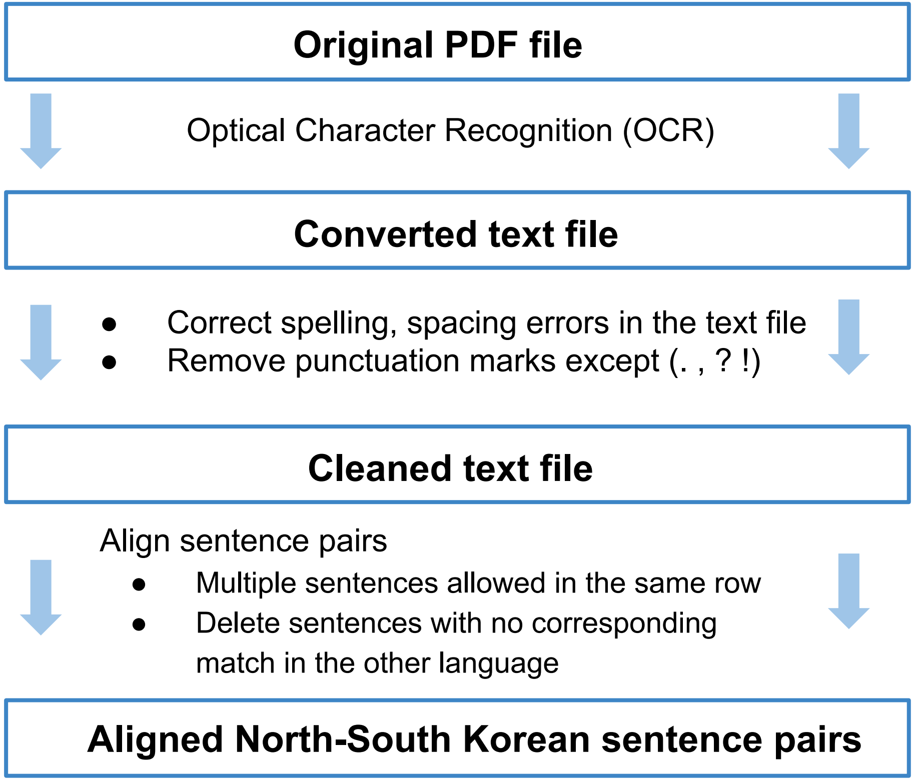
North-South Korean Parallel Corpus

North – South Korean Parallel Corpus Construction

Project Leaders



Twenty student workers



Original PDF file

Optical Character Recognition (OCR)

Converted text file

- Correct spelling, spacing errors in the text file
- Remove punctuation marks except (. , ? !)

Cleaned text file

Align sentence pairs

- Multiple sentences allowed in the same row
- Delete sentences with no corresponding match in the other language

Aligned North-South Korean sentence pairs

- Studied **linguistic differences** between the **North and South Korean** languages to **reduce errors** in the North Korean text
- Each part **reviewed twice** by different students to **reduce individual bias**

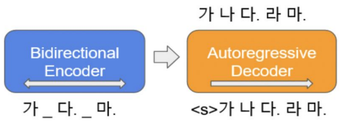
North-South Korean Parallel Corpus

North-South Korean Parallel Corpus

Resource	Title	Sentence Pairs	Total Sentence Pairs
English Classic Novel (translated version)	Jane Eyre	60,331	94,459 (72%)
French Classic Novel (translated version)	The Red and the Black	34,128	
Korean Folk Tales	Onggojip-jeon (옹고집전)	988	6,293 (5%)
	Sukhyang-jeon (숙향전)	3,538	
	Shimchung-jeon(심청전)	1,767	
Bible (translated version)	-	-	29,986 (23%)
-	-	-	130,738 (100%)

Foundation Model : KoBART

- **BART[1] (Bidirectional and Auto-Regressive Transformers)**
  - **BERT[2]**'s **bidirectional encoder** and **GPT[3]**'s **autoregressive decoder combined**
  - Denoising autoencoder that learns to map corrupted sentences to their original forms
  - Achieved high performance in various text generation task



[4]

- **KoBART[4] (South Korean BART)**
  - **BART trai**ned on approximately **40GB** of **South Korean texts** using **text infilling**
  - **Fine-tuning KoBART** with **North Korean texts** attained **high performance "North Korean-Japanese"** and **"North Korean-English"** machine translation[5]

[1] Mike Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
[2] Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
[3] Gokul Yenduri et al., "Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions", IEEE Access, 2024.
[4] https://github.com/SKT-AI/KoBART
[5] Hwichan Kim et al., "North Korean Neural Machine Translation through South Korean Resources", ACM Transactions on Asian and Low-Resource Language Information Processing, 2023.

Quantitative Analysis

- **BLEU (Bilingual Evaluation Understudy) Score**[1]

  Calculates the **surface-form similarity** between the candidate and reference texts using **n-gram overlap**, with scores ranging from 0 to 1

Model	BLEU Score
NK → SK	0.442
SK → NK	0.107

  The reason behind the higher score of the NK → SK model

  - **NK → SK** model
    - ❑ The South Korean sentences were drawn from **various publishers for each resource**
    - ❑ There were at least **one** and at most **four reference** sentence (**multiple correct answers**) to compare with the translation output
  - **SK → NK** model
    - ❑ The North Korean sentences were drawn from **one publisher for each resource**
    - ❑ There was **only one reference** sentence (**one correct answer**) to compare with the translation output

[1] Kishore Papineni et al., "Bleu: a method for automatic evaluation of machine translation", Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002.

Quantitative Analysis

- **BERT Score**[1]

  Calculates the **semantic-similarity** between candidate and reference texts using the **contextual embeddings** of BERT, with scores ranging from 0 to 1

  Effective for **paraphrase detection** and capturing **dependencies** of **unbounded length**

  - The **NK → SK** and **SK → NK model** all achieved a **high** score
  - **No significant performance difference** between the models regarding semantic similarity

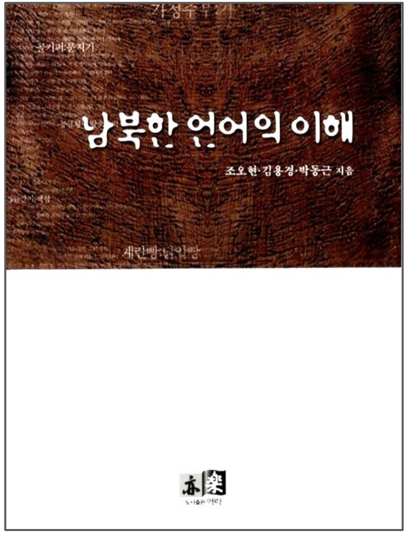
Model	BLEU Score	**BERT Score**
NK → SK	0.442	**0.821**
SK → NK	0.107	**0.815**

HAIL

[1] Tianyi Zhang et al., "BertScore: Evaluating Text Generation with BERT", 8th International Conference on Learning Representations, 2020.

Qualitative Analysis

- *Understanding the Languages of North and South Korea*[1]
  - ➢ Ch.2 : **Orthographic** differences (**spelling , word spacing**)
  - ➢ Ch.3 : **Vocabulary** difference



[1] Ohyeon Cho et al., “Understanding the Languages of North and South Korea”, Youkrack, 2002.

Qualitative Analysis

- **Vocabulary difference between North and South Korean**
    - **Loanwords**
- **Spelling difference between North and South Korean**
    - Initial sound rule
    - The addition of “ㅅ” into a compound word
    - **Endings of a word “-아/-어” based on the final syllable vowel of the stem**
- **Word spacing difference between North and South Korean**
    - **Spacing of dependent nouns and particles**
    - Spacing between main and auxiliary predicate elements

Vocabulary difference between North and South Korea

North Korea and South Korea often use **different words to refer to the same meaning**

1) Be something that everyone is interested in and talking about

『North Korean』**말밥에 오르다** (mal-bab-e o-leu-da) = 『South Korean』**화제에 오르다** (hwa-je-e o-leu-da)

Input Sentence (**NK**) →	Translation Output (**SK**)
탁자우의 나의 명함이 나의 이름을 (tag-ja-u-ui na-ui myeong-ham-i na-ui i-leum-eul) **말밥에 올리는**것이였소. (**mal-bab-e ol-li-neun**-geos-i-yeoss-so.)	테이블 위의 내 명함이 내 이름을 (te-i-beul wi-ui nae myeong-ham-i nae i-leum-eul) **화제에 올려**놓았소. (**hwa-je-e ol-lyeo**-noh-ass-so.)
A card of mine lay on the table; this being perceived, **brought my name under discussion.**	

Vocabulary difference between North and South Korea

North Korea uses **less foreign loanwords** than South Korea

1) Curtain

『North Korean』**창가림** (chang-ga-lim) = 『South Korean』**커튼** (keo-teun)

Input Sentence (**NK**) →	Translation Output (**SK**)
그래 열린 창문으로 손을 디밀어 **창가림**을 치고 (geu-lae yeol-lin chang-mun-eu-lo son-eul di-mil-eo **chang-ga-lim**-eul chi-go) 안을 들여다볼수 있을만큼 틈새를 남겨 놓았소. (an-eul deul-yeo-da-bol-su iss-eul-man-keum teum-sae-leul nam-gyeo noh-ass-so.)	열린 창문으로 손을 집어넣어 **커튼**을 젖히고 (yeol-lin chang-mun-eu-lo son-eul jib-eo-neoh-eo **keo-teun**-eul jeoj-hi-go) 안을 들여다볼 수 있을 만큼만 틈을 남겨 놓았소. (an-eul deul-yeo-da-bol su iss-eul man-keum-man teum-eul nam-gyeo noh-ass-so.)
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.	

Vocabulary difference between North and South Korea

North Korea uses **less foreign loanwords** than South Korea

1) Curtain

『North Korean』 **창가림** (chang-ga-lim) = 『South Korean』 **커튼** (keo-teun)

Input Sentence (**SK**) →	Translation Output (**NK**)
그러고는 열려 있는 창문 틈으로 손을 넣어서 창문 위로 (geu-leo-go-neun yeol-lyeo iss-neun chang-mun teum-eu-lo son-eul neoh-eo-seo chang-mun wi-lo) **커튼**을 치고 안을 살펴볼 수 있을 만큼만 공간을 남겨 두었소. (**keo-teun**-eul chi-go an-eul deul-yeo-da-bol-su iss-eul-man-keum teum-sae-leul nam-gyeo noh-ass-so.)	그리고는 열려있는 창문틈으로 손을 집어넣어 창문에 (geu-li-go-neun yeol-lyeo-iss-neun chang-mun-teum-eu-lo son-eul jib-eo-neoh-eo chang-mun-e) **창가림**을 치고 안을 들여다보게 하였다. (**chang-ga-lim**-eul chi-go an-eul deul-yeo-da-bo-ge ha-yeoss-da.)
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.	

Spelling difference between North and South Korea

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

North Korean	South Korean
When the **final syllable vowel** of the stem is “ㅣ, ㅐ, ㅔ, ㅚ, ㅟ, ㅢ” and “하” the **ending** of a word is written as **“-여/-였.”**	When the **final syllable vowel** of the stem is “ㅏ, ㅗ,” the **ending** of the word is written as **“-아.”** If not, the **ending** of the word is written as **“-어.”**
e.g., 헤여지다 (he-yeo-ji-da)	e.g., 헤어지다 (he-eo-ji-da)

Spelling difference between North and South Korea

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

1) To break off a relationship with someone
『North Korean』헤여지다 (he-yeo-ji-da) = 『South Korean』헤어지다 (he-eo-ji-da)

Input Sentence (**NK**) →	Translation Output (**SK**)
그래 나는 상당한 돈을 주고 적당한 일자리를 얻어준 (geu-lae na-neun sang-dang-han don-eul ju-go jeog-dang-han il-ja-li-leul eod-eo-jun) 다음에 체면을 유지하며 기꺼이 **헤여**졌소. (da-eum-e che-myeon-eul yu-ji-ha-myeo gi-kkeo-i **he-yeo-**jyeoss-so.)	그래서 상당한 돈을 주고 적당한 일자리를 얻어 준 (geu-lae-seo sang-dang-han don-eul ju-go jeog-dang-han il-ja-li-leul eod-eo jun) 다음부터 나는 체면을 되찾고 기꺼이 **헤어**졌소. (da-eum-bu-teo na-neun che-myeon-eul doe-chaj-go gi-kkeo-i **he-eo**-jyeoss-so.)
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.**	

Spelling difference between North and South Korea

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

1) To break off a relationship with someone
『North Korean』헤여지다 (he-yeo-ji-da) = 『South Korean』헤어지다 (he-eo-ji-da)

Input Sentence (**SK**) →	Translation Output (**NK**)
결국 충분한 돈을 주어 장사를 시작하게 해주고, (gyeol-gug chung-bun-han don-eul ju-eo jang-sa-leul si-jag-ha-ge hae-ju-go,) 깨끗이 **헤어**지고 나니 마음이 후련했소. (kkae-kkeus-i **he-eo**-ji-go na-ni ma-eum-i hu-lyeon-haess-so.)	마침내 저에게 충분한 돈을 지불하고 (ma-chim-nae jeo-e-ge chung-bun-han don-eul ji-bul-ha-go) 그와 **헤여**져 있게 되자 나는 무척 안도감을 느꼈소. (geu-wa **he-yeo**-jyeo iss-ge doe-ja na-neun mu-cheog an-do-gam-eul neu-kkyeoss-so.)
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.**	

Word Spacing difference between North and South Korea

Spacing of dependent nouns and particles

North Korean	South Korean
Dependent noun **Attach**ed to the preceding verb or adjective stem	**Dependent noun** **Separate**d from the preceding verb or adjective stem
e.g. 없을**만큼** (eobs-eul-**man-keum**)	e.g. 없을 **만큼** (eobs-eul-**man-keum**)
Particles **Attach**ed to the preceding noun	**Particles** **Attach**ed to the preceding noun
e.g. 그녀**만큼** (geu-nyeo-man-keum)	e.g. 그녀**만큼** (geu-nyeo-**man-keum**)

Word Spacing difference between North and South Korea

Spacing of dependent nouns and particles

1) Dependent noun

『North Korean』 없을**만큼** (eobs-eul-man-keum) = 『South Korean』 없을 **만큼** (eobs-eul-man-keum)

Input Sentence (**NK**) →	Translation Output (**SK**)
그 옷은 세상의 어떤 빨래하는 사람도 그보다 더 희게 (geu os-eun se-sang-ui eo-tteon ppal-lae-ha-neun sa-lam-do geu-bo-da deo hui-ge) 할수 없을**만큼** 새하얗고 눈부시게 빛났다. (hal-su eobs-eul-**man-keum** sae-ha-yah-go nun-bu-si-ge bich-nass-da.)	그 옷은 세상 어느 누구도 그보다 더 희게 (geu os-eun se-sang eo-neu nu-gu-do geu-bo-da deo hui-ge) 할 수 없을 **만큼** 새하얗고 눈부시게 빛났다. (hal su eobs-eul **man-keum** sae-ha-yah-go nun-bu-si-ge bich-nass-da.)
His clothes became dazzling white, whiter than anyone in the world could bleach them.	

HAIL

Word Spacing difference between North and South Korea

Spacing of dependent nouns and particles

2) Particle

『North Korean』그녀**만큼** (geu-nyeo-man-keum) = 『South Korean』그녀**만큼** (geu-nyeo-man-keum)

Input Sentence (**NK**) →	Translation Output (**SK**)
그 당시에 그녀**만큼** 옳바르고 흠없는 사람이 없었다. (geu dang-si-e geu-nyeo-**man-keum** olh-ba-leu-go heum-eobs-neun sa-lam-i eobs-eoss-da.)	그 당시에는 그녀**만큼** 올바르고 흠 없는 사람이 없었다. (geu dang-si-e-neun geu-nyeo-**man-keum** ol-ba-leu-go heum eobs-neun sa-lam-i eobs-eoss-da.)
She was a righteous woman, blameless among the people of her time.	

Conclusion

1. **Demonstrates the potential** of our **created North-South parallel corpus** as a valuable **resource** for developing a high-quality **machine translation model** between the North and South Korean languages
2. **The BERT Score**[1] results indicate **strong translation performance** for both the NK→SK and SK→NK models
3. **The qualitative analysis** highlights the models' ability to capture **key linguistic differences** between the North and South Korean languages

[1] Tianyi Zhang et al., "BertScore: Evaluating Text Generation with BERT", 8th International Conference on Learning Representations, 2020.

Future Work

1. Due to the **historical period** of the **literary resources** and **the Bible**, the North-South parallel corpus **lacks contemporary vocabularies** and **technical terms**
2. **Expanding the corpus** with sentence pairs from diverse sources, such as **modern literature** and **research papers from the late 20th or 21st century**, is **necessary**

Thank you for your attention

Q&A

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

- Presenter : **Hannah Hyesun Chun** (22000662@handong.ac.kr)
- Collaborators: Chanju Lee, Hyunkyoo Choi
- Advisor: Charmgil Hong (charmgil@handong.ac.kr)