Shared Heritage, Distinct Writing: Rethinking Resource Selection for East Asian Historical Documents

Seyoung Song[♢] **Haneul Yoo**[♢] **Jiho Jin**[♢] **Kyunghyun Cho**[♠♣] **Alice Oh**[♢]

[♢]KAIST [♠]New York University [♣]Genentech

{seyoung.song, haneul.yoo, jinjh0123}@kaist.ac.kr,
kyunghyun.cho@nyu.edu, alice.oh@kaist.edu

Abstract

Historical documents in the Sinosphere are known to share common formats and practices, particularly in veritable records compiled by court historians. This shared linguistic heritage has led researchers to use Classical Chinese resources for cross-lingual transfer when processing historical documents from Korea and Japan, which remain relatively low-resource. In this paper, we question the assumption of cross-lingual transferability from Classical Chinese to Hanja and Kanbun, the ancient written languages of Korea and Japan, respectively. Our experiments across machine translation, named entity recognition, and punctuation restoration tasks show minimal impact of Classical Chinese datasets on language model performance for ancient Korean documents written in Hanja, with performance differences within ± 0.0068 F1-score for sequence labeling tasks and up to $+0.84$ BLEU score for translation. These limitations persist consistently across various model sizes, architectures, and domain-specific datasets. Our analysis reveals that the benefits of Classical Chinese resources diminish rapidly as local language data increases for Hanja, while showing substantial improvements only in extremely low-resource scenarios for both Korean and Japanese historical documents. These findings emphasize the need for careful empirical validation rather than assuming benefits from indiscriminate cross-lingual transfer.

1 Introduction

Classical Chinese served as a regional lingua franca across East Asia for over a millennium, where it was used to record government chronicles, literary works, and scientific discoveries. These historical documents, particularly "veritable records" compiled by court historians, remain invaluable primary sources for studying the region's past. As Classical Chinese spread throughout East Asia, it evolved into distinct writing systems—Hanja in Korea, Kanbun in Japan, and Chữ Hán in Vietnam—collectively forming the *Sinosphere* or *Chinese character cultural sphere*.

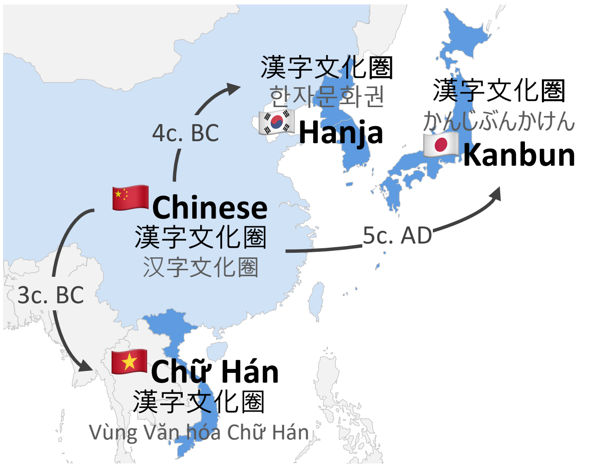


Figure 1: Language transfer from Classical Chinese to neighboring countries in Sinosphere. Classical Chinese had been transferred to neighboring countries in East Asia and used from the 6th century BC to the 20th century AD. While modern languages (*gray*) are different from each other, ancient languages (*black*) are mutually understandable.

Recent advances in natural language processing have enabled computational analysis of these historical documents, which is crucial as modern speakers can no longer directly interpret these ancient writings. Researchers are increasingly leveraging Classical Chinese resources to develop language models for other Sinosphere languages (Yoo et al., 2022; Moon et al., 2024; Wang et al., 2023, *inter alia*). This approach appears particularly promising given the shared literary traditions and significant resource disparity across these languages—with Classical Chinese being the most abundant, followed by Hanja, while Kanbun and Chữ Hán remain relatively scarce. However, the effectiveness of such cross-lingual approaches has not been thoroughly evaluated, despite these writing systems having evolved independently over 1,500 years to

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1591–1610
December 20-24, 2025 ©2025 Association for Computational Linguistics

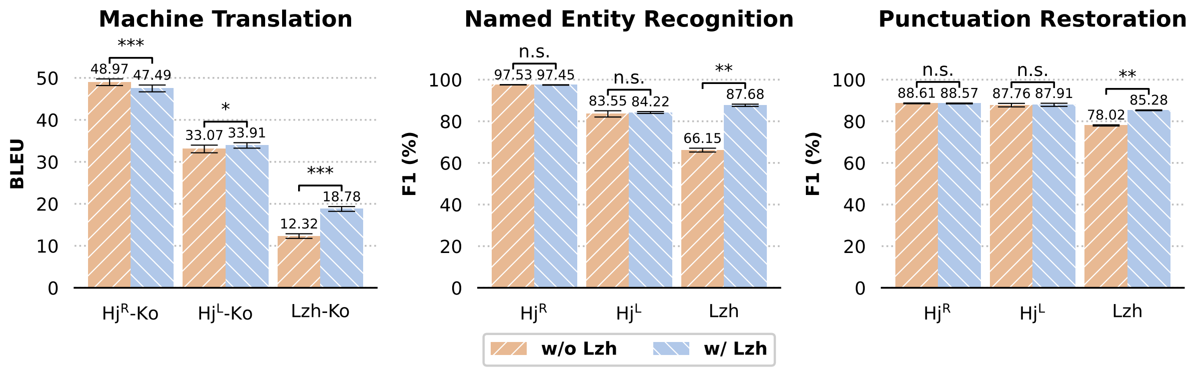


Figure 2: Comparison of models trained with and without Classical Chinese (Lzh). Results show BLEU scores (MT) and F1-scores (NER, PR) across three document types: Hanja royal records (Hj^{R}), Hanja literary works (Hj^{L}), and Classical Chinese (Lzh), with error bars of 95% confidence intervals for MT and standard deviations for NER and PR. Statistical significance is denoted as: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), and n.s. (not significant).

accommodate distinct regional needs and cultural practices.

In this paper, we challenge this assumption by conducting comprehensive experiments across three tasks: machine translation (MT), named entity recognition (NER), and punctuation restoration (PR). Figure 2 demonstrates that leveraging Classical Chinese corpora does not yield statistically significant improvements for NER and PR tasks across Hanja documents. For MT, while there is a marginally positive effect (+0.84 BLEU score) for Hanja literary works, this improvement is not substantial—according to Kocmi et al. (2024) and Xu et al. (2024), BLEU improvements of this magnitude typically correlate with human-perceived quality improvements in only 60-65% of cases. These results remain consistent across different model architectures and parameter scales, suggesting fundamental limitations in cross-lingual transfer between these historical languages (§4.1).

To enable deeper analysis beyond the predominantly royal-centric Hanja research (Kang et al., 2021; Yoo et al., 2022; Son et al., 2022, *inter alia*), we introduce *the Korean Literary Collections* (KLC), a corpus of literary works written in Hanja that captures diverse writing styles from individual scholars. Our domain-specific analysis reveals that while incorporating Classical Chinese data shows mixed results overall, careful selection of similar writing styles—such as using Chinese classical poetry for Korean literary works—can lead to marginal improvements in translation performance (§4.3).

Our investigation reveals that Classical Chinese resources provide benefit for only extremely low-resource scenarios, with their effectiveness diminishing rapidly as local language data increases for Hanja (§4.2). Experiments with Japanese historical documents written in Kanbun show similar trends of effective cross-lingual transfer in low-resource settings (§4.4.1). Moreover, our vocabulary analyses across the Sinosphere show that character-level divergence is minimal, suggesting that the limited cross-lingual transferability stems from deeper linguistic differences (§4.4.2).

Our findings across different dimensions emphasize that successful cross-lingual transfer in historical language processing requires considerations beyond shared writing systems, highlighting the importance of careful empirical validation that accounts for both resource availability and domain characteristics. Our contributions are as follows:

- We question and empirically evaluate the efficacy of leveraging Classical Chinese resources for historical Asian language models.
- We demonstrate Classical Chinese integration yields minimal improvements for Hanja processing, while showing potential benefits for extremely low-resource scenarios.
- We provide analyses of cross-lingual transfer effectiveness that can inform the development of language models for historical documents across the Sinosphere.
- We publicly release our code and data, including the KLC dataset previously unexplored in the NLP community.[1]

[1] https://github.com/seyoungsong/Shared-Heritage-Distinct-Writing

2 Background

Written languages in the Sinosphere initially adopted Classical Chinese syntax and vocabulary (Figure 1), but gradually diverged over time to meet local needs (Handel, 2019). This linguistic evolution has led to differences that potentially affect the efficacy of cross-lingual transfer in NLP tasks. First, several characters became archaic, were transformed, and substituted by preferred heteromorphic synonyms, as Classical Chinese was disseminated into neighboring countries (Kim, 2012). Table 1 illustrates examples of regional variants between languages based on Classical Chinese. Furthermore, Korea, Japan, and Vietnam developed variant forms and new characters to express local concepts (Heo, 2019). For instance, Koreans invented a new character 畓 (paddy field) in Hanja to reflect their agricultural lifestyle by combining two existing characters: 水 (water) and 田 (field). Structural adaptations also occurred; while Classical Chinese typically follows a Subject-Verb-Object (SVO) structure, Kanbun adapted to a Subject-Object-Verb (SOV) structure, aligning more closely with Japanese grammar (Wang et al., 2023).

(a) Variant forms with same meaning

Meaning	Preferred Form			
	CN	KR	JP	VN
fight	鬥	鬪	闘	鬥
truly	真	眞	真	真
leg	腳	脚	脚	腳

(b) Homographs with different meanings

Char.	Primary Meaning			
	CN	KR	JP	VN
空	in vain	empty	empty	without
骨	bone	bone	cremains	pillar
串	skewer	cape	skewer	skewer

(c) Locally invented characters

Loc.	Characters
KR	畓 (paddy field), 欌 (wardrobe)
JP	榊 (sakaki tree), 働 (work)
VN	𠀧 (three), 𠊛 (human), 𡗶 (sky)

Table 1: Linguistic divergence patterns in the Sinosphere writing systems. It illustrates three types of character variations across China (CN), Korea (KR), Japan (JP), and Vietnam (VN): variant forms sharing meanings, homographs with distinct regional interpretations, and locally invented characters.

3 Experiments

In this section, we detail the design, implementation, and results of our experiments investigating the impact of using Classical Chinese datasets to train language models for ancient Korean documents written in Hanja.

3.1 Study Design

3.1.1 Documents

We construct our dataset by gathering publicly available resources and datasets written in languages within the Sinosphere. To the best of our knowledge, resources for Kanbun and Chữ Hán are severely limited; small sizes of raw corpora exist for both, with some partial translations available for Kanbun. Therefore, we focus on Hanja (Hj) and Classical Chinese (Lzh) for our experiments. Hanja documents are further divided into two categories based on authorship: historical records written by government offices of the Joseon Dynasty (Hj^{R}) and literary work written by individual scholars (Hj^{L}). Table 2 lists these corpora with their statistics. See Appendix A for more details, including data sources and preprocessing procedures.

Royal Documents in Hanja (Hj^{R}) consists of government-compiled chronicles from the Joseon Dynasty period: *the Annals of the Joseon Dynasty* (AJD), *the Diaries of the Royal Secretariat* (DRS), and *the Daily Records of the Royal Court and Important Officials* (DRRI). These documents follow strict writing guidelines and exhibit a highly consistent style.

Literary Documents in Hanja (Hj^{L}) refers to literary works written in Hanja authored by various Korean authors. In this paper, we use *the Korean Literary Collections* (KLC) [2] as the primary source. Hanja literary works remain understudied in the NLP community, and the KLC corpus has not previously been explored in NLP research. Detailed documentation of the KLC dataset is provided in Appendix A.7.

Documents in Classical Chinese (Lzh) comprises the WYWEB benchmark (Zhou et al., 2023),

[2] also known as *the Comprehensive Publication of Korean Literary Collections in Classical Chinese*

Language	Type	Document	Time Period	Tasks			# of Samples	Avg. # of Characters	# of Tokens (GPT-4)	Trans. (%)
				MT	NER	PR				
Hanja (Hj)	Royal	AJD	1392-1928	✔	✔	✔	413,323	173.9	103,013,789	100.0
		DRS	1623-1910	✔	-	-	1,787,007	165.2	433,873,833	30.9
		DRRI	1760-1910	✔	-	-	616,910	81.1	84,141,022	32.6
	Literary	KLC	886-1933	✔	✔	✔	653,386	336.7	340,113,975	29.8
Classical Chinese (Lzh)	Mixed	Daizhige†	-	-	-	-	15,694	107,636.9	2,449,254,631	-
		NiuTrans	-	✔	-	-	972,467	22.4	31,312,241	100.0
		C2MChn†	-	✔	-	-	614,723	18.9	17,845,525	100.0
		OCDB	6 c. BC-16 c.	✔	-	-	23,795	230.9	8,018,473	100.0
		WYWMT	-	✔	-	-	266,514	21.9	8,293,026	100.0
		GLNER	-	-	✔	-	18,762	209.7	5,416,667	-
		WYWEB	1046 BC-1927	-	-	✔	135,134	117.5	22,753,344	-
Kanbun (Kb)	Royal	Rikkokushi†	697-887	✔	-	-	17,306	83.5	2,291,164	9.1
Chữ Hán	Royal	ĐVSKTT†	2 c. BC-1675	-	-	-	8,484	52.4	872,620	-
		ĐNTL†	1545-1909	-	-	-	5,608	58.8	475,523	-
		ANCL†	1285-1339	-	-	-	1,288	65.3	135,159	-
		ĐVSL†	2 c. BC-1225	-	-	-	1,164	66.3	63,677	-

Table 2: Statistics of historical documents from the Sinosphere. Documents marked with † are supplementary materials analyzed in discussions and not used in the main experimental evaluations. Trans. (%) indicates the ratio of documents with publicly available translations, and tokens are counted using tiktoken's `cl100k_base` encoding.

the NiuTrans Classical Chinese to Modern Chinese dataset[3], the C2MChn dataset (Jiang et al., 2023), Daizhige[4], and the Oriental Classics Database (OCDB)[5]. WYWEB consists of nine NLP tasks for Classical Chinese, including GLNER—a named entity recognition task initially developed by Gulian (2020)—and WYWMT—a machine translation task that translates Classical Chinese into Modern Chinese. Daizhige, the largest classical Chinese corpus, contains about 2.4 billion tokens of classical literature. The OCDB provides original Chinese texts and Korean translations of authoritative books.

Other Documents in the Sinosphere. We collect historical documents from Japan and Vietnam and analyze them in the discussion section. For Kanbun, we use the *Rikkokushi*, Japan's Six National Histories. For Chữ Hán, we include four major Vietnamese historical chronicles: the *Đại Việt sử ký toàn thư* (ĐVSKTT) and *Đại Nam thực lục* (ĐNTL), which served as official dynastic records, along with the *An Nam chí lược* (ANCL) and *Đại Việt sử lược* (ĐVSL).

Data Augmentation. We create a synthetic dataset that translates Classical Chinese into Korean by applying machine translation to Modern Chinese sentences from the NiuTrans dataset. Translation efforts for Classical Chinese predominantly focus on Modern Chinese, making it challenging to explore cross-lingual transferability. We employ GPT-4[6] to generate a total of 972,467 synthetic sentence pairs from Classical Chinese to Korean, adapting the approach proposed by Nehrdich et al. (2023). Detailed inference settings are provided in Appendix A.2.

3.1.2 Tasks

The experiments focus on three tasks: machine translation (MT), named entity recognition (NER), and punctuation restoration (PR). These tasks represent real-world challenges for human experts analyzing and understanding ancient languages.

Machine Translation (MT) of ancient Korean documents into modern languages is crucial, as most contemporary Koreans, including scholars, cannot comprehend Hanja texts without translation. We measure the BLEU score (Papineni et al., 2002) using SacreBLEU (Post, 2018).

Named Entity Recognition (NER) is a sequence labeling task that identifies and classifies proper names, such as persons and locations, in text. Combined with entity linking, it is crucial for indexing and searching large historical records. We report the F1-score after normalizing all predicted and ground-truth labels to 'NE', akin to the binary set-

[3] https://github.com/NiuTrans/Classical-Modern

[4] https://github.com/garychowcmu/daizhigev20

[5] http://db.cyberseodang.or.kr

[6] The experiments were conducted on April 6, 2024 – April 12, 2024 with `gpt-4-0125-preview` model under Azure OpenAI Service with the OpenAI API as a fallback when content filtering prevented response generation.

ting in NLTK, to ensure a fair comparison across different models and datasets. For readability, F1-scores are presented as percentages (0-100) in tables and figures, while being expressed in the standard 0-1 scale in the text (*e.g.*, 87.5 = 0.875).

Punctuation Restoration (PR) is an essential pre-translation step that involves inserting modern punctuation marks into original Hanja texts, as punctuation greatly impacts the meaning of these texts. We adopt the comprehensive punctuation restoration approach proposed by Pogoda and Walkowiak (2021) for training. For evaluation, we use the weighted average F1-score after simplifying each punctuation combination to the conventionally defined 4-class task (comma, period, question mark, and other). Reduction rules are presented in Appendix A.6.

3.1.3 Model Training

We fine-tune Qwen2-7B (Yang et al., 2024) for MT and SikuRoBERTa (Wang et al., 2021) for NER and PR, respectively. Table 8 in Appendix A.4 presents the composition of training data for each task. For documents without predefined splits, we allocate 80% for training, 10% for validation, and 10% for testing. The KLC data is bifurcated at the book level for training/validation and testing.

Qwen2 is a series of foundation models pre-trained on multilingual corpus and proficient in over 30 languages, including Chinese, Korean, and English (Yang et al., 2024). We fine-tune the Qwen2-7B using QLoRA (Dettmers et al., 2023) for machine translation of three language pairs: Hj-Ko, Hj-En, and Lzh-Ko, using the prompt in Appendix A.5.

SikuRoBERTa is a RoBERTa-based model pre-trained on the *Siku Quanshu*, a vast collection of Classical Chinese literature (Wang et al., 2021).[7]

3.2 Experimental Results

We evaluate models trained across various dataset combinations and tasks, with results shown in Table 3. Incorporating Classical Chinese resources yields minimal or non-significant improvements for Hanja documents across all tasks. For machine translation, significance testing via paired bootstrap resampling (Koehn, 2004) reveals that only 2 of 9 test conditions show improvements. The largest gain (+1.01 BLEU for Hj^{L}-Ko) achieves only 60-65% agreement with human judgments (Kocmi et al., 2024), while most conditions show decreases or stagnation (-3.14 to +0.84 BLEU). For sequence labeling tasks (*i.e.*, NER and PR), 5-fold cross-validation with Mann-Whitney U tests (Mann and

(a) Machine Translation (MT)

Train Data			**Test Data (BLEU)**			
Hj^{R}	**Hj^{L}**	**Lzh**	**Hj^{R}-En**	**Hj^{R}-Ko**	**Hj^{L}-Ko**	**Lzh-Ko**
		✔	0.02	9.79	4.85	18.13
✔			**33.16**	47.93	10.81	11.64
✔		✔	31.34	47.17	11.82	18.63
			(−1.82)	(−0.76)	(+1.01)	(+6.99)
	✔		0.13	34.16	33.57	11.91
	✔	✔	0.06	31.02	32.19	18.06
			(−0.07)	(−3.14)	(−1.38)	(+6.15)
✔	✔		33.15	**48.97**	33.07	12.32
✔	✔	✔	31.52	47.49	**33.91**	**18.78**
			(−1.63)	(−1.48)	(+0.84)	(+6.46)

(b) Named Entity Recognition (NER)

Train Data			**Test Data (F1-score)**		
Hj^{R}	**Hj^{L}**	**Lzh**	**Hj^{R}**	**Hj^{L}**	**Lzh**
		✔	81.32	72.61	86.48
✔			97.51	70.82	65.15
✔		✔	97.47	70.01	**87.85**
			(−0.04)	(−0.81)	(+22.70)
	✔		88.99	83.63	66.31
	✔	✔	86.84	83.13	87.05
			(−2.15)	(−0.50)	(+20.74)
✔	✔		**97.53**	83.55	66.15
✔	✔	✔	97.45	**84.22**	87.68
			(−0.08)	(+0.67)	(+21.53)

(c) Punctuation Restoration (PR)

Train Data			**Test Data (F1-score)**		
Hj^{R}	**Hj^{L}**	**Lzh**	**Hj^{R}**	**Hj^{L}**	**Lzh**
		✔	78.36	80.66	85.83
✔			88.58	84.77	77.25
✔		✔	88.60	84.61	85.25
			(+0.02)	(−0.16)	(+8.00)
	✔		80.49	87.05	79.45
	✔	✔	80.66	87.27	**85.95**
			(+0.17)	(+0.22)	(+6.50)
✔	✔		**88.61**	87.76	78.02
✔	✔	✔	88.57	**87.91**	85.28
			(−0.04)	(+0.15)	(+7.26)

Table 3: Performance comparisons for MT, NER, and PR tasks across all combinations of document types used in training. The values in parentheses denote the score differences between the models trained with and without Classical Chinese data (Lzh). Gray indicates no significant differences. Orange and blue indicate significant decreases and increases, respectively, with saturation reflecting the magnitude of differences by each task. **Bold** and underlined numbers denote the highest and the second-highest scores for each task and test dataset, respectively.

[7] Encoder-based models pretrained on Classical Chinese corpora have been employed by multiple Hanja-related studies (Yoo et al., 2022; Moon et al., 2024).

Whitney, 1947) shows no significant changes ($p < 0.05$) when adding Classical Chinese data, with F1-score differences ranging from -0.0215 to +0.0067. In contrast, Classical Chinese documents show significant performance improvements when trained with Classical Chinese resources, indicating successful baseline training. A qualitative error analysis of these results is available in Appendix B.6.

Notably, models trained exclusively on Classical Chinese perform well on sequence labeling tasks for Hanja documents, with the Classical Chinese NER model outperforming Hj^{R}-trained model on Hj^{L} data (0.7261 vs 0.7082 F1). While machine translation requires comprehensive language understanding and generation capabilities, NER and PR primarily capture character and word-level patterns. The smaller performance variations in PR compared to MT and NER suggest that punctuation patterns exhibit a degree of consistency across the Sinosphere writing systems.

Our results reveal a clear division between royal and literary Hanja texts. Models trained on Hj^{R} perform poorly on Hj^{L} (BLEU scores below 11.82), with similar patterns in NER. This aligns with known linguistic differences between government chronicles, which follow strict guidelines, and diverse literary works by individual authors (Moon et al., 2024).

For Classical Chinese language modeling, incorporating Hanja data shows minimal impact. Adding Hj^{L} produces no significant changes across tasks, while Hj^{R} data yields modest differences (+0.50 BLEU, +0.0137 F1, and -0.0058 F1 for MT, NER, and PR, respectively).

4 Discussions

In this section, we explore potential reasons why Classical Chinese exhibits limited impact on the language models for Asian historical documents and support them with empirical analyses.

4.1 Model Scaling and Architecture Variations

We extend our observations to smaller model scales (Table 4) and various foundation models (Table 5) by fine-tuning MT models with and without Classical Chinese data. We outline that incorporating Classical Chinese corpora significantly impairs Hanja language modeling across both smaller scales of Qwen2 and different foundation models (*i.e.*, Llama-3.1-8B-Instruct and Gemma-2-9B). Specifically, BLEU scores for Hanja-to-English and Hanja-to-Korean on royal documents decrease by 5.08 and 5.94, respectively, when fine-tuning Qwen2-1.5B.

Model Size	Train Hj	Train Lzh	Hj^{R}-En	Hj^{R}-Ko	Hj^{L}-Ko	Lzh-Ko
7B	✔		**33.15**	**48.97**	33.07	12.32
	✔	✔	31.52	47.49	**33.91**	**18.78**
			(−1.63)	(−1.48)	(+0.84)	(+6.46)
1.5B	✔		28.74	43.58	29.32	8.92
	✔	✔	23.66	37.64	26.66	15.61
			(−5.08)	(−5.94)	(−2.66)	(+6.69)
0.5B	✔		17.34	34.14	21.30	3.45
	✔	✔	14.38	33.01	16.77	10.17
			(−2.96)	(−1.13)	(−4.53)	(+6.72)

Table 4: BLEU scores of machine translation models at varying parameter scales trained with/without Classical Chinese (Lzh) data.

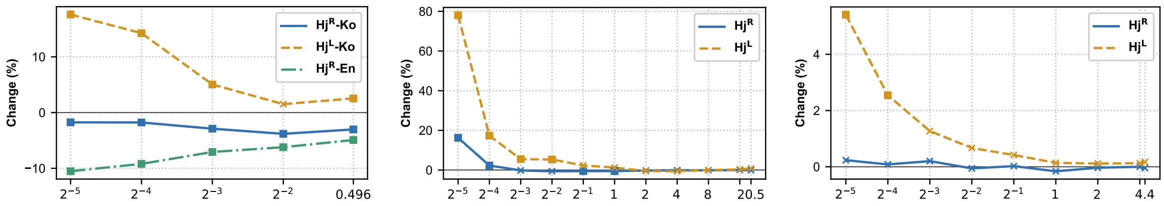
Model	Train Hj	Train Lzh	Hj^{R}-En	Hj^{R}-Ko	Hj^{L}-Ko	Lzh-Ko
Qwen2	✔		33.15	48.97	33.07	12.32
	✔	✔	31.52	47.49	33.91	18.78
			(−1.63)	(−1.48)	(+0.84)	(+6.46)
Llama-3.1	✔		33.96	49.03	34.56	13.13
	✔	✔	32.25	47.53	33.50	18.76
			(−1.71)	(−1.50)	(−1.06)	(+5.63)
Gemma-2	✔		**35.39**	**51.86**	**36.69**	13.20
	✔	✔	33.56	49.66	35.09	**19.61**
			(−1.83)	(−2.20)	(−1.60)	(+6.41)

Table 5: BLEU scores of machine translation models across different architectures with/without Classical Chinese (Lzh) training data.

4.2 Threshold for Diminishing Benefits of Classical Chinese Data

We hypothesize that sufficient Hanja data exists to train effective language models without relying on Classical Chinese resources, given the substantial volume of annotated Hanja documents preserved through national research initiatives. When measured by token count, available training data for Hanja exceeds Classical Chinese by factors of 4.4, 18.6, and 6.8 for MT, NER, and PR, respectively.

To identify the threshold where Classical Chinese data ceases to provide meaningful benefits, we conduct an ablation study by systematically varying the ratio of Hanja to Classical Chinese training data. Figure 3 shows performance differences between models trained with and without Classical Chinese data across different Hanja data proportions. While Classical Chinese resources significantly boost performance in extremely low-resource scenarios, particularly for literary documents, these



(a) Machine Translation

(b) Named Entity Recognition

(c) Punctuation Restoration

Figure 3: Performance impact of Classical Chinese training data across varying Hanja data ratios. The x-axis shows the ratio r, where Hj:Lzh = r:1 denotes the proportion of Hanja data against Classical Chinese data, while the y-axis shows the relative performance differences in percentage (%) between models trained with/without Classical Chinese data. Square and x markers indicate statistically significant differences ($p < 0.05$) and non-significant differences, respectively.

benefits diminish rapidly as Hanja data increases. The performance improvements become relatively small (below 5.5%) across all tasks once Hanja data exceeds one-eighth the volume of Classical Chinese data. Detailed results are in Table 15. These findings suggest that while Classical Chinese resources can be valuable in low-resource settings, their utility diminishes quickly with increasing Hanja data availability, challenging the assumption that incorporating additional auxiliary data consistently improves performance.

4.3 Domain-Specific Transfer Learning

We further investigate whether targeting specific domains of Classical Chinese data can improve cross-lingual transfer effectiveness for Hanja. Using the C2MChn dataset (Jiang et al., 2023), we categorize Classical Chinese texts into three domains aligned with Hanja genres: History, Religion (Buddhism, Confucianism, Taoism), and Miscellaneous (Agronomy, Short, Others), and conduct fine-tuning experiments with Qwen2-7B using various domain combinations.

Table 6 shows that incorporating Classical Chinese data from any domain combination reduces MT model performance for Hanja royal documents compared to using Hanja data alone. While the Miscellaneous domain occasionally produces minor improvements for literary documents (maximum +1.41 BLEU), the overall effects remain mixed or negligible. We hypothesize that short-form poetry within the Miscellaneous domain may assist with similarly styled Hanja literary works, but using untargeted data across domains diminishes this benefit. These results underscore that domain-specific Classical Chinese data requires careful empirical validation for effective use.

Domain			Hj^R-En	Hj^R-Ko	Hj^L-Ko	Lzh-Ko
His	Rel	Mis				
None (baseline)			**33.15**	**48.97**	33.07	12.32
✔			32.26 (−0.89)	47.80 (−1.17)	33.60 (+0.53)	16.88 (+4.56)
	✔		32.23 (−0.92)	47.82 (−1.15)	33.68 (+0.61)	16.90 (+4.58)
		✔	<u>32.71</u> (−0.44)	<u>48.55</u> (−0.42)	**34.48** (+1.41)	16.78 (+4.46)
✔	✔		31.98 (−1.17)	47.97 (−1.00)	32.27 (−0.80)	**17.52** (+5.20)
✔		✔	31.89 (−1.26)	47.45 (−1.52)	34.03 (+0.96)	16.83 (+4.51)
	✔	✔	31.80 (−1.35)	48.11 (−0.86)	<u>34.06</u> (+0.99)	16.96 (+4.64)
✔	✔	✔	31.77 (−1.38)	47.37 (−1.60)	33.66 (+0.59)	<u>17.47</u> (+5.15)

Table 6: Performance comparison of domain-specific transfer learning for machine translation. Models are trained on Hanja data (351.1M tokens) combined with different domains of Classical Chinese: History (23.6M tokens), Religion (21.6M tokens), and Miscellaneous (3.7M tokens).

4.4 Expandability to Sinosphere

4.4.1 Machine Translation for Kanbun

To explore the generalizability of our findings to other languages in the Sinosphere, we conduct experiments on Kanbun using 1,371 paragraph-level samples from Korean-related records [8] in the Six National Histories of Japan. As shown in Table 7, both Hanja and Classical Chinese resources improve Kanbun translation performance (BLEU scores increase by 19.17 and 11.14, respectively), demonstrating that cross-lingual transfer can be effective in low-resource settings. However, careful empirical validation is needed when selecting source languages rather than simply combining all

[8] https://db.history.go.kr/id/jm

Train Data			Kb-Ko	Hj^{R}-Ko	Hj^{L}-Ko	Lzh-Ko
Kb	Hj	Lzh				
✔			25.96	8.02	4.50	10.29
	✔		13.82	48.97	33.07	12.32
		✔	19.08	9.79	4.85	18.13
✔	✔		**45.13**	**49.53**	**34.69**	14.00
✔		✔	37.10	9.70	4.85	17.88
	✔	✔	19.14	47.49	33.91	**18.78**
✔	✔	✔	42.66	47.93	33.69	18.40

Table 7: Translation performance (BLEU score) comparison across different combinations of Kanbun (Kb, 0.34M tokens), Hanja (351.1M tokens), and Classical Chinese (79.8M tokens) training data. The **bold** and underlined values indicate the best and second-best performance, respectively.

available resources.

Here, the varying degrees of improvement likely stem from different levels of linguistic and topical similarity. We validate this empirically using 5-gram language models trained on Korean translations, where perplexity on Kanbun documents is lower with a model trained on Hanja (181) versus Classical Chinese (264). This pattern reflects our test set composition: Korea-related Kanbun texts translated by a Korean institution.

4.4.2 Vocabulary Divergence

We computationally identify the linguistic distance between Classical Chinese and other writing systems in the Sinosphere through character-level analysis. Analysis of unique characters across writing systems (Figure 4) reveals Hanja having the largest vocabulary (23,186 characters), followed by Classical Chinese, Chữ Hán, and Kanbun. While 32.2% of Hanja characters do not appear in our Classical Chinese corpus, these Hanja-exclusive characters occur infrequently, comprising less than 1.9% of character usage at the 99% frequency threshold (Figure 5). Further inspection reveals that most Hanja-exclusive characters are documented variant forms of Classical Chinese characters in the *Kangxi Dictionary*, rather than Korean-invented characters. For instance, the character 腦 in the Annals of the Joseon Dynasty is a known variant of 腦 (brain) but absent from our Classical Chinese corpora. While variant character normalization techniques (Kessler, 2024) might mitigate these surface-level differences, our findings suggest that the challenges in cross-lingual transfer stem from factors beyond vocabulary divergence.

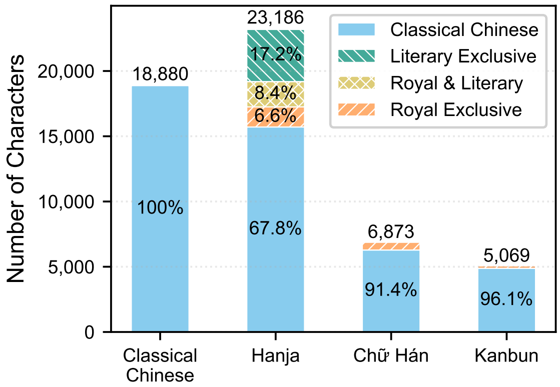


Figure 4: Distribution of unique characters across writing systems in the Sinosphere. The bars represent the proportion of shared characters with Classical Chinese versus language-specific variants in each writing system.

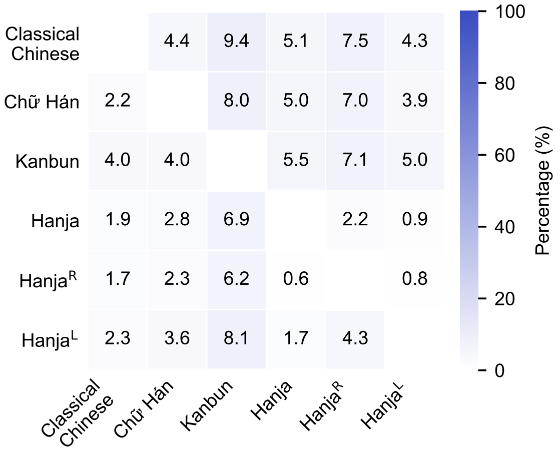


Figure 5: Heatmap of character coverage gaps between Sinosphere languages. Each cell shows the percentage of characters in the *row* language that are not the most common characters in the *column* language at 99% frequency threshold.

5 Related Work

5.1 NLP for Asian Historical Documents

A variety of research has been mainly conducted in Classical Chinese and Hanja due to challenges for acquisition of available resources. In Classical Chinese, evaluation datasets and benchmarks (Zhou et al., 2023) and language models (Tian et al., 2021; Chang et al., 2023) are widely released. Similarly, datasets and language models for Hanja have been introduced for various tasks, including machine translation (Kang et al., 2021; Son et al., 2022), named entity recognition (Yoo et al., 2022), and relation extraction (Yang et al., 2023).

5.2 Cross-Lingual Studies for Sinosphere

Several studies have introduced cross-lingual approaches that leverage linguistically close, historical resources in the Sinosphere. Moon et al. (2024) used Classical Chinese resources to develop NER and sentence splitting models for Hanja literary documents and uncovered that removing special characters and punctuation marks helps cross-lingual transfer between Classical Chinese and Hanja. Wang et al. (2023) synthetically constructed the first Classical Chinese-to-Kanbun dataset and trained a Kanbun language model, addressing the scarcity of available resources in Kanbun.

Cross-lingual transfer in the Sinosphere has also been explored across modern languages. Kim et al. (2020) proposed a machine translation technique that matches overlapping vocabulary between Korean and Japanese stemming from Hanja and Kanbun, respectively. Nehrdich et al. (2023) used Classical Chinese-to-Modern Chinese dataset for Buddhist Chinese-to-English machine translation. While recent studies have recklessly adopted Classical Chinese resources for other languages in the Sinosphere, this paper aims to carefully investigate the performance of cross-lingual transfer.

6 Conclusion

We challenge the widespread assumption that Classical Chinese resources inherently benefit language models for other historical East Asian writing systems. Our comprehensive experiments across machine translation, named entity recognition, and punctuation restoration reveal that incorporating Classical Chinese data produces minimal and often statistically insignificant improvements for Hanja documents. While our analysis shows limited character-level divergence between these languages, the poor cross-lingual transfer suggests fundamental linguistic differences beyond shared vocabulary. These findings demonstrate that successfully processing historical Asian languages requires careful empirical validation rather than assumed benefits from apparent linguistic similarities. We emphasize the importance of considering both resource availability and domain characteristics when developing language models for historical documents. Building on our results, future research should further investigate the linguistic factors that affect cross-lingual transferability across different languages or writing systems.

Limitations

Our experiments with Kanbun and Chữ Hán are constrained by limited dataset availability compared to Hanja, necessitating caution in drawing broader conclusions about these writing systems. Also, as NLP researchers rather than domain experts in historical Asian languages, our analysis may not fully capture deeper linguistic nuances in ancient languages.

Despite analyzing substantial volumes of historical records and literary work, our coverage of Hanja documents remains partial. Notable omissions include local government records, Buddhist texts, and epigraphic sources, which may demonstrate distinct patterns of cross-lingual transferability from Classical Chinese.

The representation of Classical Chinese texts in our datasets poses an additional limitation, as they are available only in Simplified Chinese despite their Traditional Chinese origins. This inherently imperfect character conversion system may introduce systematic biases in our cross-lingual analysis.

Ethical Considerations

This research focuses on evaluating the effectiveness of cross-lingual transfer between historical writing systems through computational experiments on publicly available historical documents. The methods employed are applied to texts that have been openly preserved for academic study. The research does not involve human subjects, sensitive personal data, or content that could enable harmful applications. While historical texts can sometimes contain biased perspectives or sensitive content, our work focuses purely on the technical aspects of language processing rather than interpreting or generating content. The computational methods and findings presented here aim to advance the scholarly study of historical documents while maintaining respect for the cultural significance of these texts.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00509258 and No. RS-2024-00469482, Global AI Frontier Lab).

This work was also supported by the Samsung Advanced Institute of Technology (under the project

Next Generation Deep Learning: From Pattern Recognition to AI).

This research project has benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program through which leading foundation models hosted by Microsoft Azure along with access to Azure credits were provided to conduct the research.

We acknowledge using Claude[9] and ChatGPT[10] for writing and coding assistance.

References

Liu Chang, Wang Dongbo, Zhao Zhixiao, Hu Die, Wu Mengcheng, Lin Litao, Shen Si, Li Bin, Liu Jiangfeng, Zhang Hai, and Zhao Lianzheng. 2023. SikuGPT: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *Preprint*, arXiv:2304.07778.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Gulian. 2020. "Gulian Cup" Ancient Book Document Named Entity Recognition Competition of CCL 2020.

Zev Handel. 2019. *Sinography: The Borrowing and Adaptation of the Chinese Script*. Brill, Leiden, The Netherlands.

Chul Heo. 2019. From the point of view of academic terms, the term ’han gukgoyuhanja (韓國固有漢字)’ is proposed as a way to solve the problem of classification and name of ’han-character system’. *The Oriental Studies*, 75:147–164.

Zongyuan Jiang, Jiapeng Wang, Jiahuan Cao, Xue Gao, and Lianwen Jin. 2023. Towards better translations from classical to modern chinese: A new dataset and a new method. In *Natural Language Processing and Chinese Computing: 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12–15, 2023, Proceedings, Part I*, pages 387–399, Berlin, Heidelberg. Springer-Verlag.

Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4031–4042, Online. Association for Computational Linguistics.

Florian Kessler. 2024. Towards context-aware normalization of variant characters in classical Chinese using parallel editions and BERT. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 141–151, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Eunhee Kim. 2012. 한자의 수용과 변용: 한자의 특성과 중국 남방 漢字系文字의 제자원리. *중국언어연구*, 41:173–203.

Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi. 2020. Korean-to-Japanese neural machine translation system using hanja information. In *Proceedings of the 7th Workshop on Asian Translation*, pages 127–134, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP ’23, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Henry B Mann and Donald R Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Hyeonseok Moon, Myunghoon Kang, Jaehyung Seo, Sugyeong Eo, Chanjun Park, Yeongwook Yang, and Heuiseok Lim. 2024. Exploiting hanja-based resources in processing korean historic documents written by common literati. *IEEE Access*, 12:59909–59919.

Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody, and Kurt Keutzer. 2023. MITRA-zh: An efficient, open machine translation solution for buddhist

[9] https://claude.ai

[10] https://chatgpt.com

Chinese. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 266–277, Tokyo, Japan. Association for Computational Linguistics.

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.

Michał Pogoda and Tomasz Walkowiak. 2021. Comprehensive punctuation restoration for English and Polish. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4610–4619, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary Korean and English. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1260–1272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, and Bin Li. 2021. SikuBERT SikuRoBERTa：面向字人文的《四全》模型建及用究. *Library Tribune*.

Hao Wang, Hirofumi Shimizu, and Daisuke Kawahara. 2023. Kanbun-LM: Reading and translating classical Chinese in Japanese methods by language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8589–8601, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada. Association for Computational Linguistics.

Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. HUE: Pretrained model and dataset for understanding hanja documents of Ancient Korea. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1832–1844, Seattle, United States. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319, Toronto, Canada. Association for Computational Linguistics.

Appendix

A Replication Details

A.1 Data Sources

We collect our datasets from publicly available sources between February and October 2024. Korean historical documents are sourced from national research institutions: the National Institute of Korean History (NIKH) provides the AJD[11] and DRS[12], while the Kyujanggak Institute maintains DRRI[13]. The Institute for the Translation of Korean Classics (ITKC) offers the KLC[14] along with Korean translations of the royal documents. Classical Chinese resources include Daizhige[15], NiuTrans[16], C2MChn[17], and WYWEB[18], all available through GitHub repositories. The OCDB[19] is maintained by the Institute of Traditional Culture. For Japanese documents, we use the Rikkokushi texts from the public website[20], with Korean translations of Korea-related records provided by NIKH[21]. Vietnamese historical chronicles including ĐVSKTT[22], ĐNTL[23], ANCL[24], and ĐVSL[25] are available through Wikisource.

A.2 Data Augmentation

We create synthetic Korean translations of Classical Chinese texts using GPT-4. For each source text, we provide both the Classical Chinese original and its Modern Chinese translation as context, using the following prompt:

> Translate the following text from Classical Chinese into Korean, based on the reference translation in Modern Chinese. ↵
> Classical Chinese: <source sentence> ↵
> Modern Chinese: <reference translation> ↵
> Korean:

We generate translations using GPT-4 under two configurations: the NiuTrans dataset translations use `gpt-4-0125-preview` with temperature 0.7, while C2MChn translations use `gpt-4o-mini-2024-07-18` with temperature 0.0. We employ Azure OpenAI Service as our primary platform, falling back to the OpenAI API when necessary. Approximately 6% of source texts are filtered out due to sensitive historical content, particularly passages containing references to war crimes or violence.

A.3 Preprocessing

Processing ancient Asian texts requires careful character normalization to ensure consistent representation across different writing systems and time periods. Our preprocessing pipeline applies the Normalization Form Compatibility Composition (NFKC) to standardize character encodings, followed by whitespace standardization that converts all newlines, tabs, and spaces to single space characters. We normalize all punctuation marks, including converting directional quotation marks to their neutral forms, and standardize CJK middle dot variants (`U+318D`, `U+119E`, `U+30FB`) to the standard middle dot form (`U+00B7`). For Classical Chinese texts in Simplified Chinese characters, we convert them to Traditional Chinese using OpenCC[26].

A.4 Experimental Setup

Table 8 quantifies our experimental data across tasks using both sample counts and token quantities. Table 9 presents our dataset partitioning across training, validation, and test sets for each task. For machine translation (MT), we evaluate performance using 1,000 test samples per document and language pair, computing aggregate BLEU scores

Task	Type	Document	# of Samples	# of Tokens
MT	Hj$^{\text{R}}$	AJD	331,150	241,653,871
	Hj$^{\text{L}}$	KLC	53,147	109,406,346
	Lzh	NiuTrans	774,914	79,806,362
NER	Hj$^{\text{R}}$	AJD	293,854	80,841,316
	Hj$^{\text{L}}$	KLC	8,035	6,673,763
	Lzh	GLNER	14,719	4,710,310
PR	Hj$^{\text{R}}$	AJD	293,746	81,095,372
	Hj$^{\text{L}}$	KLC	14,428	7,983,038
	Lzh	WYWEB	70,664	13,141,862

Table 8: Composition of training data used in experiments across tasks. Data quantities are shown by both number of samples and total tokens computed using `cl100k_base` encoding.

[11] https://sillok.history.go.kr

[12] https://sjw.history.go.kr

[13] https://kyudb.snu.ac.kr/series/main.do?item_cd=ILS

[14] https://db.itkc.or.kr

[15] https://github.com/garychowcmu/daizhigev20

[16] https://github.com/NiuTrans/Classical-Modern

[17] https://github.com/Zongyuan-Jiang/C2MChn

[18] https://github.com/baudzhou/WYWEB

[19] https://db.cyberseodang.or.kr

[20] http://www.kikuchi2.com/sheet/rikkoku.html

[21] https://db.history.go.kr/id/jm

[22] https://zh.wikisource.org/wiki/大越史記全書

[23] https://zh.wikisource.org/wiki/大南寔錄

[24] https://zh.wikisource.org/wiki/安南志

[25] https://zh.wikisource.org/wiki/越史略

[26] https://github.com/BYVoid/OpenCC

Tasks	Type	Document	Lang.	Train	Val	Test
MT	Hj^{R}	AJD	Hj-En	16,032	0	1,000
			Hj-Ko	299,106	0	1,000
			Ko-En	16,012	0	1,000
		DRS	Hj-Ko	0	0	1,000
		DRRI	Hj-Ko	0	0	1,000
	Hj^{L}	KLC	Hj-Ko	53,147	0	1,000
	Lzh	NiuTrans	Lzh-Ko	774,914	0	1,000
		WYWMT	Lzh-Ko	0	0	1,000
		OCDB	Lzh-Ko	0	0	1,000
		C2MChn†	Lzh-Ko	542,305	0	0
	Kb	Rikkokushi†	Kb-Ko	1,025	0	346
NER	Hj^{R}	AJD	Hj	293,854	37,830	5,000
	Hj^{L}	KLC	Hj	8,035	995	5,000
	Lzh	GLNER	Lzh	14,719	2,000	2,000
PR	Hj^{R}	AJD	Hj	293,746	37,831	5,000
	Hj^{L}	KLC	Hj	14,428	1,797	5,000
	Lzh	WYWEB	Lzh	70,664	32,607	5,000

Table 9: Dataset composition and partitioning across tasks. The table shows sample sizes for training, validation, and test sets used in machine translation (MT), named entity recognition (NER), and punctuation restoration (PR) experiments. Documents marked with † are supplementary materials used only in discussions.

via SacreBLEU across all translation outputs. For named entity recognition (NER) and punctuation restoration (PR), we use 5,000 test samples per document, with the exception of GLNER, which uses 2,000 test samples due to dataset constraints.

A.5 Training and Hyperparameters

Our experiments run on a server equipped with Intel Xeon Silver 4114 processor (40 threads) and eight GeForce RTX 2080 Ti GPUs (11GB each). For training and inference of Gemma-2 models, we use a separate server with Intel Xeon Silver 4214R processor (48 threads) and eight Quadro RTX A6000 GPUs (48GB each). We implement our models using LLaMA-Factory (Zheng et al., 2024) for machine translation fine-tuning and Hugging Face Transformers (Wolf et al., 2020) for NER and PR models. Table 10 details our hyperparameter configurations. Training times vary by task: up to 36 hours for machine translation, 10 hours for named entity recognition, and 14 hours for punctuation restoration. The prompt shown below is used consistently across all translation tasks during both training and inference.

```
Translate the following text from <source language> into
<target language>. ⏎
<source language>: <source sentence> ⏎
<target language>:
```

Hyperparameter	Value
Max sequence length	512
Batch size	64
Initial checkpoint	`Qwen/Qwen2-7B`
Quantization	4-bit NormalFloat and double quantization
LoRA r	16
LoRA α	32
LoRA dropout	0.0
rsLoRA	True
Number of epochs	1
Learning rate	1.0e-4
Learning rate scheduler	Cosine
Warm-up ratio	0.1
Optimizer	8-bit AdamW
Weight decay	0.01
Gradient clipping	1.0

(a) Hyperparameters for MT models.

Hyperparameter	Value
Max sequence length	512
Batch size	32
Initial checkpoint	`SIKU-BERT/sikuroberta`
Max epochs	5
Early stopping	applied on validation loss
Learning rate	2e-4
Learning rate scheduler	Linear
Warm-up ratio	0.1
Optimizer	AdamW
Weight decay	0.01

(b) Hyperparameters for NER and PR models.

Table 10: Hyperparameter configurations for training MT, NER, and PR models. Values shown for MT models use `Qwen/Qwen2-7B` base architecture (additional experiments use `Qwen/Qwen2-1.5B`, `Qwen/Qwen2-0.5B`, `google/gemma-2-9b`, and `meta-llama/Llama-3.1-8B-Instruct`). We use half precision (`fp16`) for all computation.

A.6 Inference and Evaluation

Machine Translation. We quantize the fine-tuned MT models using AWQ (Lin et al., 2024) and utilize vLLM (Kwon et al., 2023) for inference. The prompt used for training is also used for inference. We set the temperature to 0 and employ greedy decoding. Metric signatures and versions used for evaluation are presented in Table 12.

Punctuation Restoration. For evaluation, we simplify the diverse punctuation marks used in the original documents and our models into a standardized 4-class scheme consisting of COMMA, PERIOD, QUESTION, and OTHER. This allows for consistent comparison of model performance across the different datasets. Table 13 shows how various punctuation characters are mapped to these four classes based on their typical functions or meanings.

A.7 Korean Literary Collections Dataset

For this study, we compile a new dataset from the Korean Literary Collections (KLC), a comprehensive collection of Hanja literary works maintained by the Institute for the Translation of Korean Classics. Unlike prior research that focused predominantly on royal-centric Hanja documents (Kang et al., 2021; Yoo et al., 2022; Son et al., 2022), our KLC dataset captures diverse writing styles from individual scholars spanning from 886 to 1933 CE, with particularly rich coverage during the 1800s-1930s period. The source corpus contains 652,622 unique articles with an average length of 337 Hanja characters (approximately 220M characters total) from 1,258 unique authors, including notable historical figures such as Song Si-yeol (宋時烈), Jeong Yak-yong (丁若鏞), and Kwak Jong-seok (郭鍾錫). Table 11 presents the genre distribution of the translated portion, demonstrating substantial coverage beyond official documents. We structured our KLC dataset to support multiple NLP tasks: raw text for language model pretraining (652,622 samples), parallel data for machine translation (157,202 samples with Hanja-Korean translations), and annotated documents for named entity recognition (21,657 samples with 379,976 entities).

Genre	# of Articles	Ratio (%)
Anthology	112,215	71.4
Miscellaneous	11,707	7.4
Travelogue	9,688	6.2
Literature	5,456	3.5
History	4,035	2.6
Complete Collection	3,878	2.5
Law	1,609	1.0
Ceremonial Texts	1,593	1.0
Human Affairs	1,422	0.9
Astronomy	1,075	0.7
Politics	944	0.6
Medicine	779	0.5
Geography	664	0.4
Agriculture	618	0.4
Philosophy	595	0.4
Ceremonial Records	452	0.3
Foreign Relations	364	0.2
Classical Texts	53	0.0
Mathematics	41	0.0
Archives	14	0.0

Table 11: Genre distribution of the translated portion in the Korean Literary Collections (KLC) dataset, showing the number of articles and percentage for each category.

B Complementary Results

This section presents additional experimental results and analyses that complement our main findings.

B.1 Experimental Results

Table 14 provides comprehensive BLEU scores for machine translation experiments across all dataset combinations and language pairs, including results from different model architectures and training configurations.

B.2 Threshold for Diminishing Benefits

Table 15 details our systematic investigation of how varying the ratio between Hanja and Classical Chinese training data affects model performance. The results encompass performance metrics across machine translation, named entity recognition, and punctuation restoration tasks as we gradually reduce the proportion of Hanja data on a logarithmic scale.

B.3 Machine Translation for Kanbun

Figure 6 illustrates how BLEU scores change as the quantity of additional training data decreases for Kanbun-Korean translation. The relative performance advantages between different systems remain consistent across varying data quantities.

B.4 Vocabulary Divergence

Figure 7 presents the proportion of unique characters in each corpus that do not appear in other corpora, measured at four cumulative frequency thresholds: 100%, 99.9%, 99%, and 95%. This analysis reveals the extent of character-level divergence between writing systems in the Sinosphere.

B.5 Analysis of Performance in Low-Resource Settings

To verify that the benefits observed when adding Classical Chinese resources in low-resource scenar-

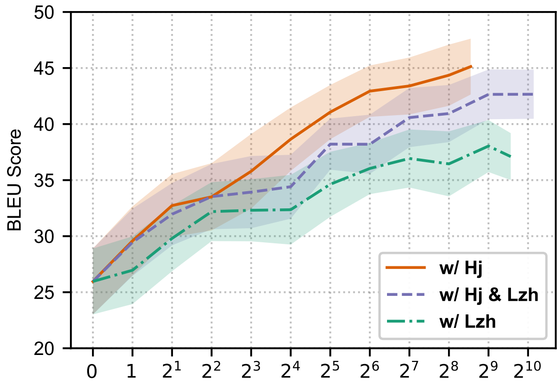
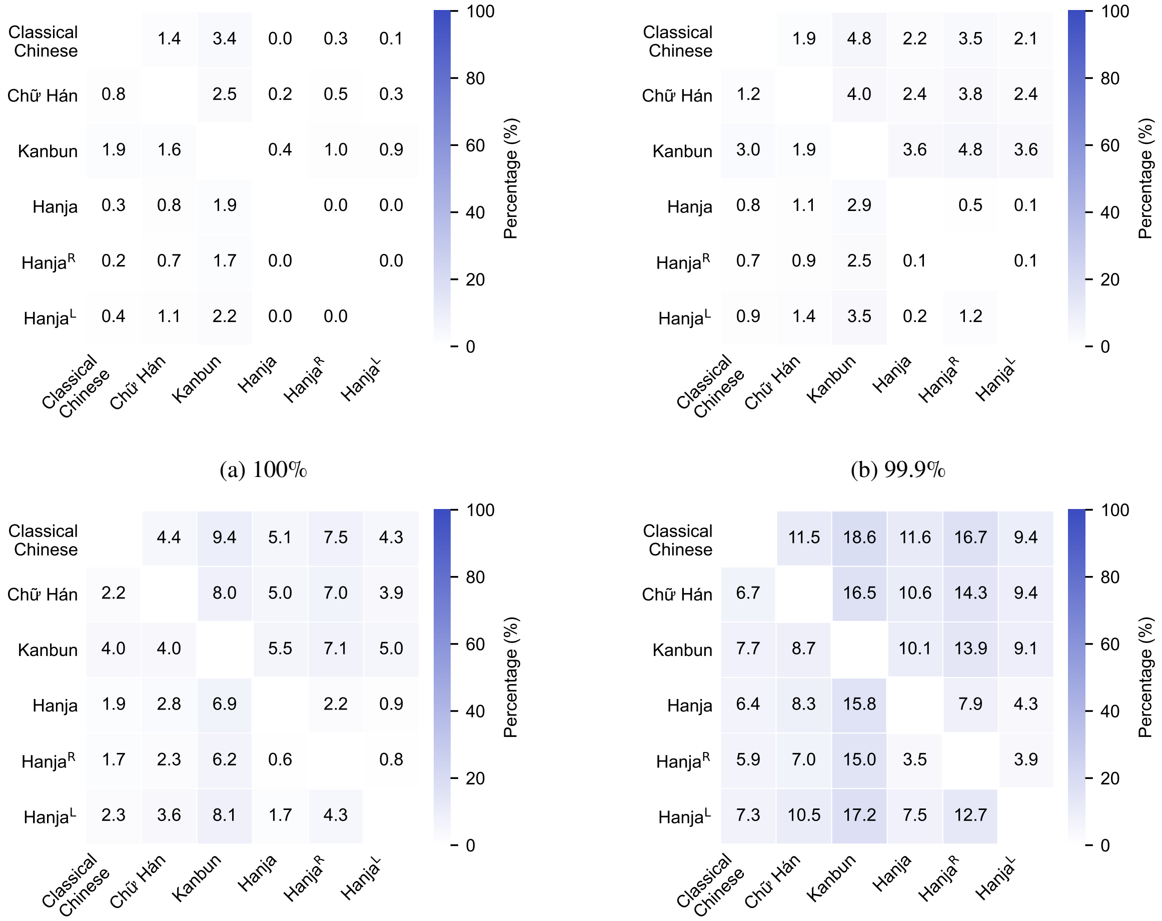


Figure 6: Performance comparison of Kanbun-Korean translation models with varying amounts of additional training data. The x-axis shows the ratio of additional data to Kanbun data in $\log_2$ scale, and the y-axis shows BLEU scores with 95% confidence intervals indicated by shaded regions.

ios reflect genuine cross-lingual transfer rather than overfitting mitigation, we conduct additional experiments analyzing evaluation loss behavior. We maintain full validation set sizes while systematically reducing training data for Hanja-only models at two extreme low-resource ratios (1/16 and 1/32 of the original data). Figure 8 shows that evaluation loss decreases monotonically across all settings for both NER and PR tasks, with no indication of validation loss increases or plateauing that would typically signal overfitting. This consistent pattern across different data ratios strongly suggests that models trained on extremely limited Hanja data do not suffer from overfitting, even without Classical Chinese data. Therefore, the performance improvements observed when adding Classical Chinese resources in these settings likely represent genuine benefits from cross-lingual transfer rather than simply regularization effects addressing overfitting issues.

B.6 Qualitative Error Analysis

To complement our quantitative findings, we perform a systematic qualitative analysis comparing outputs of models trained with and without Classical Chinese data. We calculate per-sample performance metrics for all test predictions and categorize instances where the inclusion of Classical Chinese resources leads to performance changes. Table 16 presents representative examples from our analysis of Hanja^{R} to Korean translation, which reveals three recurring error patterns: (1) inappropriate modernization of classical terms, where historically specific terminology is simplified into contemporary equivalents (*e.g.*, “찬구(饌具)” → “반찬”, replacing a formal historical term for food provisions with a modern casual word for side dishes); (2) loss of Korea-specific concepts, where terms unique to Korean historical and cultural contexts are omitted or generalized (*e.g.*, “황의장(黃儀仗)” → “의장”, losing the Korea-specific royal ceremonial context); and (3) name translation errors, where historical Korean names are inconsistently handled (*e.g.*, “윤방(尹滂)” → “윤팽”, incorrectly changing the pronunciation). These patterns suggest that Classical Chinese data can introduce biases that obscure culturally and historically specific nuances in Hanja translation, explaining the quantitative performance degradation observed in §3.2. For sequence labeling tasks (NER and PR), our analysis shows no consistent patterns of improvement or degradation, aligning with the statistical non-significance reported in our main results.



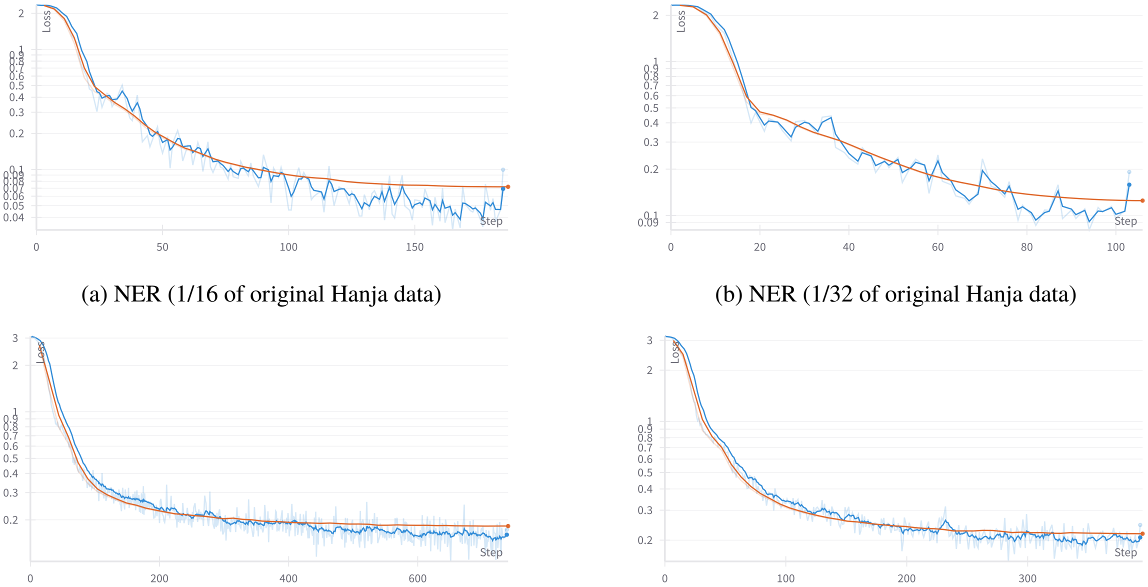
(a) 100%

(b) 99.9%

(c) 99%

(d) 95%

Figure 7: Character divergence patterns across writing systems at different frequency thresholds.



(a) NER (1/16 of original Hanja data)

(b) NER (1/32 of original Hanja data)

(c) PR (1/16 of original Hanja data)

(d) PR (1/32 of original Hanja data)

Figure 8: Evaluation loss curves for NER and PR tasks in low-resource settings. Blue lines and orange lines represent training loss and validation loss, respectively.

<table>
<tr><th>Metric</th><th>Version</th></tr>
<tr><td>BLEU [En]</td><td>nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2</td></tr>
<tr><td>BLEU [En] Paired-bootstrap resampling</td><td>nrefs:1|bs:2000|seed:42|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2</td></tr>
<tr><td>BLEU [Ko]</td><td>nrefs:1|case:mixed|eff:no|tok:ko-mecab-0.996/ko-0.9.2-KO|smooth:exp|version:2.4.2</td></tr>
<tr><td>BLEU [Ko] Paired-bootstrap resampling</td><td>nrefs:1|bs:2000|seed:42|case:mixed|eff:no|tok:ko-mecab-0.996/ko-0.9.2-KO|smooth:exp|version:2.4.2</td></tr>
<tr><td>BLEU [Zh]</td><td>nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.4.2</td></tr>
<tr><td>BLEU [Zh] Paired-bootstrap resampling</td><td>nrefs:1|bs:2000|seed:42|case:mixed|eff:no|tok:zh|smooth:exp|version:2.4.2</td></tr>
</table>

Table 12: Metric versions and signatures.

<table>
<tr><th>Class</th><th>Characters</th></tr>
<tr><td>COMMA</td><td>- (U+002D), / (U+002F), : (U+003A), | (U+007C), · (U+00B7), 、 (U+3001)</td></tr>
<tr><td>PERIOD</td><td>! (U+0021), . (U+002E), ; (U+003B), 。 (U+3002)</td></tr>
<tr><td>QUESTION</td><td>? (U+003F)</td></tr>
</table>

Table 13: Punctuation reduction rules for simplifying diverse punctuation marks in the punctuation restoration task to a standardized 4-class scheme: COMMA, PERIOD, QUESTION, and OTHER.

Model	Train Data							Test Data (BLEU)										
	Hj^R	Hj^L	Lzh				Kb	Hj^R				Hj^L	Lzh					Kb
	AJD	KLC	Niu	C2MChn			Rikko-	AJD		DRS	DRRI	KLC	OCDB	NiuTrans		WYWMT		Rikkokushi
			Trans	His	Rel	Mis	kushi	Hj-En	Hj-Ko	Hj-Ko	Hj-Ko	Hj-Ko	Lzh-Ko	Lzh-Ko	Lzh-Zh	Lzh-Ko	Lzh-Zh	Kb-Ko
Qwen2-7B	-	-	✔	-	-	-	-	0.02	10.96	10.35	7.22	4.85	12.93	26.25	5.75	21.60	6.18	19.08
	✔	-	-	-	-	-	-	33.16	55.13	47.39	39.64	10.81	14.63	9.13	20.70	7.26	13.38	-
	✔	-	✔	-	-	-	-	31.34	52.49	46.40	39.03	11.82	13.71	26.65	18.58	21.62	14.02	-
	-	✔	-	-	-	-	-	0.13	38.34	34.67	28.22	33.57	14.11	9.88	20.22	8.53	10.73	-
	-	✔	✔	-	-	-	-	0.06	35.59	30.22	26.11	32.19	12.94	26.12	10.51	21.57	8.66	-
	✔	✔	-	-	-	-	-	33.15	55.30	48.65	40.65	33.07	16.13	9.42	15.13	7.33	8.74	13.82
	✔	✔	✔	-	-	-	-	31.52	52.83	47.04	39.33	33.91	14.26	26.06	1.21	21.68	0.86	19.14
Qwen2-1.5B	✔	✔	-	-	-	-	-	28.74	50.69	43.32	35.02	29.32	11.12	7.66	1.78	5.42	0.92	-
	✔	✔	✔	-	-	-	-	23.66	45.58	36.02	29.89	26.66	11.03	23.14	0.11	18.30	0.05	-
Qwen2-0.5B	✔	✔	-	-	-	-	-	17.34	43.34	31.20	27.08	21.30	2.90	4.75	1.84	3.64	1.02	3.79
	✔	✔	✔	-	-	-	-	14.38	41.55	30.90	25.16	16.77	5.13	19.15	0.20	13.81	0.18	-
Gemma-2-9B	✔	✔	-	-	-	-	-	35.39	58.24	52.15	43.14	36.69	16.40	9.76	2.63	9.02	2.57	-
	✔	✔	✔	-	-	-	-	33.56	55.89	49.45	41.48	35.09	14.69	27.60	0.06	22.68	0.07	-
Llama-3.1-8B-Instruct	✔	✔	-	-	-	-	-	33.96	56.00	48.67	40.45	34.56	16.78	9.31	6.57	8.90	6.48	-
	✔	✔	✔	-	-	-	-	32.25	54.21	47.05	39.26	33.50	14.00	26.24	18.65	21.93	12.62	-
Qwen2-7B	✔	✔	-	✔	-	-	-	32.26	54.02	47.65	39.44	33.60	15.02	20.06	4.88	17.99	4.03	-
	✔	✔	-	-	✔	-	-	32.23	53.26	47.40	39.42	33.68	16.12	18.95	9.71	16.62	6.44	-
	✔	✔	-	-	-	✔	-	32.71	54.94	47.48	40.70	34.48	16.06	18.71	10.97	16.56	8.17	-
	✔	✔	-	✔	✔	-	-	31.98	53.62	47.82	39.39	32.27	15.75	20.95	5.95	18.16	4.13	-
	✔	✔	-	✔	-	✔	-	31.89	54.39	46.46	39.40	34.03	14.75	20.72	3.74	17.73	3.16	-
	✔	✔	-	-	✔	✔	-	31.80	54.01	47.65	40.11	34.06	16.04	19.29	6.14	16.78	4.90	-
	✔	✔	-	✔	✔	✔	-	31.77	52.86	47.39	38.68	33.66	15.79	20.83	9.50	18.03	6.77	-
Qwen2-7B	-	-	-	-	-	-	✔	7.50	8.56	8.43	6.58	4.50	10.51	10.66	22.17	9.46	16.57	25.96
	✔	✔	-	-	-	-	✔	33.32	55.23	49.30	41.29	34.69	17.78	10.04	20.11	9.13	11.13	45.13
	-	-	✔	-	-	-	✔	0.02	10.62	10.66	6.93	4.85	12.72	25.73	1.70	21.49	1.76	37.10
	✔	✔	✔	-	-	-	✔	31.31	51.45	48.57	39.05	33.69	13.29	26.35	6.17	21.95	5.56	42.66

Table 14: Comprehensive BLEU scores for machine translation experiments.

Train Data Ratio (Hj : Lzh)	Hj^{R}				Hj^{L}	Lzh				
	AJD		DRS	DRRI	KLC	OCDB	NiuTrans		WYWMT	
	Hj-En	Hj-Ko	Hj-Ko	Hj-Ko	Hj-Ko	Lzh-Ko	Lzh-Ko	Lzh-Zh	Lzh-Ko	Lzh-Zh
0.496 : 0	33.15	55.30	48.65	40.65	33.07	16.13	9.42	15.13	7.33	8.74
0.496 : 1	31.52	52.83	47.04	39.33	33.91	14.26	26.06	1.21	21.68	0.86
2^{-2} : 0	31.26	52.01	47.15	39.21	31.80	15.72	9.93	20.47	8.45	11.81
2^{-2} : 1	29.32	51.29	45.37	37.54	32.28	14.18	25.69	8.30	22.09	7.53
2^{-3} : 0	29.00	51.01	45.42	36.02	29.15	14.68	9.15	19.75	7.55	11.73
2^{-3} : 1	26.95	48.38	42.75	36.83	30.62	12.94	26.13	10.78	21.66	10.09
2^{-4} : 0	26.63	47.25	39.72	33.36	25.35	12.91	8.42	22.64	7.06	14.67
2^{-4} : 1	24.18	47.51	37.13	34.01	28.96	13.71	25.92	8.38	22.20	9.05
2^{-5} : 0	23.20	43.70	37.25	30.97	23.76	11.52	8.35	26.19	7.28	18.17
2^{-5} : 1	20.76	44.76	35.37	29.93	27.94	13.28	26.05	4.10	21.88	4.46
0 : 0	-	-	-	-	-	-	-	-	-	-
0 : 1	0.02	10.96	10.35	7.22	4.85	12.93	26.25	5.75	21.60	6.18

(a) MT (BLEU)

Train Data Ratio (Hj : Lzh)	Hj^{R}	Hj^{L}	Lzh
	AJD	KLC	GLNER
20.5 : 0	97.53	83.55	66.15
20.5 : 1	97.45	84.22	87.68
2^{4} : 0	97.39	83.42	65.92
2^{4} : 1	97.40	83.71	87.83
2^{3} : 0	97.14	82.41	65.82
2^{3} : 1	97.00	82.39	87.77
2^{2} : 0	96.63	80.94	65.28
2^{2} : 1	96.53	80.43	87.54
2^{1} : 0	96.07	78.70	64.83
2^{1} : 1	95.81	78.30	87.20
1 : 0	95.33	76.25	64.03
1 : 1	94.81	77.19	87.06
2^{-1} : 0	94.26	72.48	62.37
2^{-1} : 1	93.74	74.16	86.83
2^{-2} : 0	92.94	68.82	60.48
2^{-2} : 1	92.35	72.46	86.83
2^{-3} : 0	90.44	65.54	56.76
2^{-3} : 1	90.26	69.15	86.58
2^{-4} : 0	85.64	62.31	52.14
2^{-4} : 1	87.58	73.10	86.69
2^{-5} : 0	73.97	41.18	34.32
2^{-5} : 1	85.99	73.31	86.60
0 : 0	-	-	-
0 : 1	81.32	72.61	86.48

(b) NER (F1)

Train Data Ratio (Hj : Lzh)	Hj^{R}	Hj^{L}	Lzh
	AJD	KLC	WYWEB
4.36 : 0	88.61	87.76	78.02
4.36 : 1	88.57	87.91	85.28
2^{2} : 0	88.54	87.74	78.12
2^{2} : 1	88.54	87.85	85.42
2^{1} : 0	87.99	87.17	77.89
2^{1} : 1	87.96	87.27	85.76
1 : 0	87.39	86.65	77.62
1 : 1	87.25	86.77	85.76
2^{-1} : 0	86.65	86.00	77.35
2^{-1} : 1	86.67	86.36	85.84
2^{-2} : 0	85.95	85.28	76.95
2^{-2} : 1	85.90	85.85	85.88
2^{-3} : 0	84.93	84.19	76.31
2^{-3} : 1	85.10	85.26	85.93
2^{-4} : 0	83.60	82.20	74.87
2^{-4} : 1	83.67	84.29	85.92
2^{-5} : 0	81.16	79.17	72.89
2^{-5} : 1	81.35	83.45	85.87
0 : 0	-	-	-
0 : 1	78.36	80.66	85.83

(c) PR (F1)

Table 15: Ablation study results showing model performance across varying ratios of Hanja (Hj) to Classical Chinese (Lzh) training data for (a) machine translation measured by BLEU score, (b) named entity recognition measured by F1 score, and (c) punctuation restoration measured by F1 score. Ratios range from using only Lzh data (0:1) to the full Hj:Lzh ratio for each task. † denotes evaluation on augmented data.

Error Type	Source Hanja Text	Ground-truth Korean Translation	Prediction (Model w/o Lzh)	Prediction (Model w/ Lzh)
Modernization of classical terms	諭黃海道觀察使曰: “延安付處李瓔, 依住家舍, 令所在邑隨宜修理, 當身及奴婢幷五口, 月給糧料、酒肉、饌具。”	황해도 관찰사(黃海道觀察使)에게 유시(諭示)하기를, “연안(延安)에 부처(付處)한 이영(李瓔)이 의지해 살 가옥을 소재한 고을로 하여금 적절히 수리하게 하고, 그 자신과 노비(奴婢) 5구(口)에게 아울러 매월 양료(糧料) · 주육(酒肉) · 찬구(饌具)를 지급하라.” 하였다.	황해도 관찰사(黃海道觀察使)에게 유시(諭示)하기를, “연안(延安)에 부처(付處)한 이영(李瓔)은 그의 거처하는 집을 소재읍(所在邑)에서 적당히 수리하게 하고, 그 자신과 노비(奴婢) 아울러 5구(口)에게는 달마다 양료(糧料)와 술 · 고기 · 찬구(饌具)를 주도록 하라.” 하였다.	황해도 관찰사에게 유시하기를, “연안(延安)에 부처(付處)한 이영(李瓔)은 그가 사는 집에 의거하여 거처하게 하고, 소재(所在)하는 고을로 하여금 적당히 수리하게 하며, 당신과 노비와 아울러 5명에게는 매월 양식 · 술 · 고기 · 반찬을 주도록 하라.” 하였다.
Loss of Korea-specific concepts	政院啓曰: “經理都監郎廳來言: ’自中原, 表裏出來于經理處云, 故經理明日欲迎於郊外, 黃儀仗取來’ 云。”	정원이 아뢰었다. “경리 도감 낭청이 와서 ‘중국에서 표리(表裏)를 경리에게 보내왔다 하므로 경리가 내일 교외에서 맞이하려고 하는데 황의장(黃儀仗)을 가져 오라고 했다.’ 하였습니다.”	정원이 아뢰었다. “경리 도감 낭청이 와서 말하기를 ‘중원에서 표리(表裏)가 경리에게 나왔다고 하므로 경리가 내일 교외에서 맞이하려고 하니 황의장(黃儀仗)을 가져오라.’고 하였습니다.”	정원이 아뢰었다. “경리 도감 낭청이 와서 말하기를 ‘중원에서 표리(表裏)가 나와 경리처에 왔다.’고 하므로 경리가 내일 교외에서 맞이하고자 하니, 의장(儀仗)을 가져오라고 하였다.”
Name translation errors	臺諫啓前事, 命遞崔連孫、尹滂、成希仲、閔慶安, 餘不允。	대간이 전의 일을 아뢰니 최연손 · 윤방 · 성희중 · 민경안은 체직하도록 하고, 나머지는 윤허하지 않았다.	대간이 전의 일을 아뢰니, 최연손 · 윤방 · 성희중 · 민경안은 체직시키고 나머지는 윤허하지 않았다.	대간이 전의 일을 아뢰니, 최연손 · 윤팽 · 성희중 · 민경안은 체직시키라 명하고 나머지는 윤허하지 않았다.

Table 16: Examples of translation errors when incorporating Classical Chinese resources