Open Korean Historical Corpus: A Millennia-Scale Diachronic Collection of Public Domain Texts

Seyoung Song[1] **Nawon Kim**[2] **Songeun Chae**[1] **Kiwoong Park**[1]
Jiho Jin[1] **Haneul Yoo**[1] **Kyunghyun Cho**[3] **Alice Oh**[1]
[1]KAIST [2]Korea University [3]New York University
seyoung.song@kaist.ac.kr, knawon08@korea.ac.kr,
{songeun, marspak, jinjh0123, haneul.yoo}@kaist.ac.kr,
kyunghyun.cho@nyu.edu, alice.oh@kaist.edu

Abstract

The history of the Korean language is characterized by a discrepancy between its spoken and written forms and a pivotal shift from Chinese characters to the Hangul alphabet. However, this linguistic evolution has remained largely unexplored in NLP due to a lack of accessible historical corpora. To address this gap, we introduce the Open Korean Historical Corpus, a large-scale, openly licensed dataset spanning 1,300 years and 6 languages, as well as under-represented writing systems like Korean-style Sinitic (Idu) and Hanja-Hangul mixed script. This corpus contains 17.7 million documents and 5.1 billion tokens from 19 sources, ranging from the 7th century to 2025. We leverage this resource to quantitatively analyze major linguistic shifts: (1) Idu usage peaked in the 1860s before declining sharply; (2) the transition from Hanja to Hangul was a rapid transformation starting around 1890; and (3) North Korea's lexical divergence causes modern tokenizers to produce up to 51 times higher out-of-vocabulary rates. This work provides a foundational resource for quantitative diachronic analysis by capturing the history of the Korean language. Moreover, it can serve as a pre-training corpus for large language models, potentially improving their understanding of Sino-Korean vocabulary in modern Hangul as well as archaic writing systems.

Keywords: Korean, Hanja, Diachronic corpus, Historical linguistics, Mixed script, Low-resource languages

1. Introduction

Written communication on the Korean peninsula before the 20th century is linguistically compelling due to the inherent discrepancy between its spoken language and writing systems. Prior to the creation of the native alphabet, Hangul (한글), Koreans borrowed Hanja (漢字) and used Classical Chinese as the primary literary language. Because the grammatical structure of Korean (SOV) differs from that of Classical Chinese (SVO), unique transcription systems like Idu were developed to better reflect native grammar (Handel, 2019). Even after the invention of Hangul in the 15th century, a centuries-long transition involving mixed script usage preceded the predominantly Hangul-based writing of today. The very success of this transition, however, has rendered centuries of Hanja-based records illegible to most contemporary Koreans, making its accessibility a significant challenge for modern NLP.

However, computational research on Korean has predominantly focused on modern, Hangul-based texts, largely neglecting the intermediate period of Hanja-Hangul mixed script. A significant barrier to this research is the lack of accessible historical corpora. Most corpora created by the Korean government (*e.g.*, Sejong Corpus, Modu Corpus) are distributed under restrictive licenses that make releasing derivative works nearly impossible (Hwang and Choi, 2016). While some institutions release digitized documents online, they often do not offer them as downloadable datasets. As a result, interested researchers have been forced to individually crawl numerous websites to compile their own datasets.

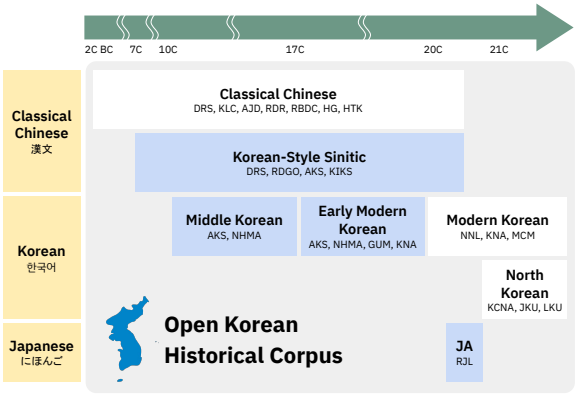

Figure 1: Overview of Open Korean Historical Corpus, which covers major languages and distinct writing systems used on the Korean peninsula. The abbreviations for each source are explained in Table 2. The blue cells indicate the corpora that we organized and released for the first time under open license terms.

To address this problem, we introduce the Open Korean Historical Corpus, a large-scale, openly licensed dataset. We collected and processed 17.7

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6561–6572
11–16 May 2026.

Period	Language	Text
19c–	**Modern Korean** (Hangul)	그래도 벌써 몇 년 전 일입니다. But it was already several years ago.
20c–	**North Korean** (Hangul)	닭 료리 는 먹지 않겠어. 남새료리 를 먹자. I won't eat chicken dishes. Let's eat vegetable dishes.
19c–	**Modern Korean** (Hanja-Hangul Mixed Script)	標 로 풀게된 漢文易解法 Chart-Based Method for Interpreting Classical Chinese
19–20c	**Japanese** (Colonial Era)	又大ニ民心 ヲ 收 ムル ニ似 タリ It also seemed that they greatly won the hearts of the people.
17–19c	**Early Modern Korean** (Hanja-Old Hangul Mixed Script)	驚怯逃走 하야 不得禁防 이올두 相考施行伏望 Frightened, they fled, and prevention was impossible. We respectfully request your examination and action.
17–19c	**Early Modern Korean** (Old Hangul)	밧바 이만 뎍노라 I'm busy, so I'll leave it at this.
10–16c	**Middle Korean** (Old Hangul)	엇디 인ᄂᆞᆫ다 먼 ᄃᆡ 긔별 도 요ᄉᆞ이ᄂᆞᆫ 더옥 듣 디 몯ᄒᆞ니 념녜 무궁 ᄒᆞ다 How are you? As I cannot hear news from afar these days, my worries are endless.
7–20c	**Korean-style Sinitic** (Idu)	此明文內乙用良告官辨正爲乎事 The matter is hereby submitted to the authorities for review and determination.
–20c	**Classical Chinese** (Hanja)	八月庚辰 月暈, 內赤中黃, 又與歲星, 同舍. On Gyeongjin day in the 8th month, a red and yellow lunar halo appeared with Jupiter.

Table 1: Textual diversity in the Open Korean Historical Corpus. Selected texts illustrate the linguistic diversity across different eras (periods are approximate). Highlighting distinguishes key scripts and lexical features—such as Hanja, Japanese, Old Hangul, and North Korean vocabulary—from standard Hangul. An English translation is provided for each example.

million documents from 19 distinct sources, totaling 5.1 billion tokens and covering the period from the 7th century through 2025. As illustrated in Figure 1, the collection encompasses six languages: Korean (Middle, Early Modern, Modern, North), Classical Chinese, and Modern Japanese. This is, to our knowledge, the first openly licensed corpus to offer broad temporal coverage of these historical stages, particularly Middle Korean and Early Modern Korean, as well as under-represented writing systems, including Korean-style Sinitic (Idu) and Hanja-Hangul mixed script. Table 1 provides selected examples that illustrate this linguistic diversity across different eras.

Using this corpus, we quantitatively analyze major linguistic shifts. Our analysis of Korean-style Sinitic (Idu) shows its usage peaked in the 1860s before declining sharply, particularly after the 1894 Kabo Reform. We also trace the diachronic transition from Hanja to Hangul, finding it was a rapid transformation rather than a gradual shift; writing exclusively in Classical Chinese dominated written records until 1890, but by 1980, Hangul comprised over 93% of characters. Furthermore, we quantify the lexical divergence of North Korean, which causes modern tokenizers to produce up to 51 times higher out-of-vocabulary (OOV) rates due to its unique orthography for loanwords and distinct native vocabulary. We release the corpus and our processing code under the CC BY-NC 4.0 and MIT licenses, respectively, to facilitate further research in Korean diachronic linguistics and historical NLP.[1]

2. Background

In this section, we examine the historical background of the Korean language and writing system and discuss how they have changed over time.

2.1. Historical Development of Korean

The Korean language, with its earliest extant records dating back to the early 5th century (Kim, 1998; Bailblé, 2016), has undergone substantial changes. While multiple periodization schemes have been proposed, reflecting different scholarly foci and periodization criteria, we adopt the four-period scheme (NIKL, 2010) for consistency and clarity. This provides a useful overview by combining dynastic transitions and major linguistic changes as criteria for division.

Ancient (–10C) This period spans from before the Common Era to the establishment of the Goryeo dynasty in 918. During this period, Chinese characters were introduced to the Korean peninsula and became the primary medium of written Korean, known as *Hanja* (Choo, 2016). Only a few written records from this era have survived.

[1]Dataset and code available at https://github.com/seyoungsong/OKHC.

Medieval (10C–16C) The creation of *Hangul*, the native Korean alphabet, in 1443 marked a turning point in the history of the Korean language (Pae, 2018). Hangul aimed to mitigate the disparity between native spoken Korean and written Korean borrowed from Chinese characters. Yet its impact was gradual, as written Korean continued to rely heavily on *Hanja*. At the same time, the language differed from its modern form, most notably through a tonal system.

Early Modern (17C–late 19C) The Korean language underwent significant changes around the 17th century. In particular, it experienced a gradual loss of its tonal system, including the tonal mark (*pangjeom*, 傍點, "side marks") and the vowel (*arae-a*, ㆍ). In addition, the overall grammar structure became simpler than in earlier periods. Furthermore, the devastation following the Imjin War (1592–1598) led to the loss of older texts and the weakening of the orthographic conventions that had been maintained in previous eras.

Modern (late 19C–present) During this period, Korean underwent dynamic changes shaped by major historical events. In the late 19th century, Hangul was granted official status as the national script in the wake of the modernization movement. However, Japanese colonial rule (1910–1945) restricted the use of Korean language and taught Japanese ideology and language (Pak and Hwang, 2011). Afterwards, it gradually diverged into distinct varieties in the north and the south, following the division of the peninsula (1945).

2.2. Diverse Writing Systems in Korean

Korean has undergone complex, multi-layered changes in its writing systems.

Korean-style Sinitic After the presumed introduction of Chinese characters around the 2nd century BCE (Eom, 2002), the disparity between spoken Korean and written Chinese motivated the Korean-style Sinitic systems, such as *Idu* (writing Chinese characters in Korean word order) (King, 2022), *Gugyeol* (adding grammatical markers to Classical Chinese texts) (Kim, 2004), and *Hyangchal* (used mainly for transcribing vernacular poetry, *hyangga*) (Lee and Ramsey, 2011).

Sino-Korean Mixed Script This system refers to the combined use of *Hanja* and *Hangul*, and reflects the coexistence of the two writing systems following the invention of *Hangul* in the 15th century (Pae, 2018). This system became widespread in the late 19th century and continued into the modern era (Crellin and Joyce, 2019).

> 第九條, 法律命令은 다 國文으로써 本을 삼ᄋᆞᆫ 漢譯을 附ᄒᆞ며 或國、漢文을 混用홈。
>
> *Article 9: All laws and ordinances shall use the national script as the standard, accompanied by a Chinese translation, or may be written in a mixed script.*

Orthographic Divergence after Korean Division While both North and South Korea established their orthographic systems based on the Unified Korean Orthography (1933), they have diverged through separate revision processes after the division (Lee, 2021). Moreover, in South Korea, standardization was primarily based on the Seoul dialect, whereas in North Korea, it was centered on the Pyongyang variety (Song, 2015).

3. Open Korean Historical Corpus

This section details the construction of the Open Korean Historical Corpus, outlining the methodology for data collection, text preprocessing, language identification, and schema design. We provide a statistical overview of the corpus and a discussion of the legal considerations for its release.

3.1. Data Collection

To build a comprehensive resource, we target key languages and writing systems with historical significance on the Korean peninsula. The languages include several varieties of Korean (Middle, Early Modern, Modern, and North Korean), Classical Chinese, and Japanese. The inclusion of Japanese addresses its enforced use during the colonial period (1910–1945), which left a substantial body of written records and influenced the Korean language post-independence. The corpus covers a range of writing systems, including Korean-style Sinitic (Idu), Hanja-Hangul mixed script, Old Hangul, modern Hangul, and the Japanese writing system. Sources are selected based on their ability to provide digitized original texts accompanied by publication year metadata. Data is gathered primarily by web scraping institutional websites using Python libraries such as BeautifulSoup, HTTPX, and Selenium, and by direct downloads where available. Data collection occurred from May 2025 to October 2025, with the most recent documents from North Korean news sources dated June 19, 2025.

Attribution. The corpus is composed of materials from numerous institutions, as detailed in Table 2. A significant portion of the historical documents is sourced from the National Institute of

Abb.	Source	Lic.	Pub. Years	Size	Documents	Avg. Len.	Languages
NNL	Naver News Library	▲	1920-1999	13 GB	13,536,494	385	Modern Korean
DRS	Diaries of the Royal Secretariat	✓	1623-1910	750 MB	1,792,187	165	CC, Idu
KLC	Korean Literary Collections	✓	886-1933	656 MB	652,405	335	CC
KCNA	Korean Central News Agency	✗	1998-2025	320 MB	170,472	741	North Korean
KNA	Korean Newspaper Archive	✓	1883-1952	187 MB	364,409	210	Modern Korean
AJD	Annals of the Joseon Dynasty	✓	1392-1928	182 MB	413,131	173	CC
RDR	The Records of Daily Reflections	✓	1760-1910	152 MB	338,084	153	CC
RDGO	Records and Documents of the Government Offices	✓	1637-1910	130 MB	130,143	380	CC, Modern Korean, Idu
MCM	Modern and Contemporary Magazines	✓	1896-1943	121 MB	15,326	3,228	Modern Korean
GUM	GongU Madang	✓	1019-1995	119 MB	13,291	3,725	Modern Korean, CC
JKU	Journal of Kim Il Sung University	✗	2014-2025	104 MB	39,723	1,261	North Korean
AKS	Academy of Korean Studies	✓	695-1985	72 MB	55,482	502	CC, Early Mod Ko, Mid Ko, Idu
RBDC	Records of the Border Defense Council	✓	1616-1892	55 MB	93,528	233	CC
RJL	Records of the Japanese Legation	✓	1893-1913	35 MB	22,502	617	Japanese
KIKS	Kyujanggak Institute for Korean Studies	✓	1395-1953	23 MB	32,487	263	CC, Idu
HG	History of Goryeo	✓	1451	4 MB	20,047	79	CC
NHMA	National Hangeul Museum Archive	✓	1628-1988	3 MB	1,669	743	Modern Korean, Early Mod Ko
LKU	Literary Works of Kim Il Sung University	✗	2015-2025	1 MB	274	1,759	North Korean
HTK	History of the Three Kingdoms	✓	1145	667 KB	4,613	58	CC
Total			**695-2025**	**16 GB**	**17,696,267**	**357**	

Table 2: Overview of the data sources and statistics. The Lic. column indicates accessibility: ✓ for public domain or openly licensed; ▲ for partially restricted sources where texts published before 1963 are public domain; and ✗ for copyrighted materials, for which only metadata and links are provided. Avg. Len. denotes the average document length in characters. In the Languages column, CC, Early Mod Ko, and Mid Ko refer to Classical Chinese, Early Modern Korean and Middle Korean, respectively. Idu, technically a writing system rather than a distinct language, is also listed in this column.

Korean History (NIKH)[2], which provides the *Diaries of the Royal Secretariat*[3], the *Annals of the Joseon Dynasty*[4], *Records and Documents of the Government Offices*, *Modern and Contemporary Magazines*, the *Records of the Border Defense Council*, the *Records of the Japanese Legation*, the *History of Goryeo*, and the *History of the Three Kingdoms*. Additional public archives include the National Library of Korea (NLK)[5], which contributes the *Korean Newspaper Archive*, and the National Hangeul Museum[6], which supplies its *Archive*. The Institute for the Translation of Korean Classics (ITKC)[7] provides the *Korean Literary Collections*, while the Kyujanggak Institute for Korean Studies[8] offers *The Records of Daily Reflections* and its collection of old documents. The Academy of Korean Studies (AKS)[9] contributes a range of materials, including old Korean books, document collections, royal court documents, and Hangul letters, which we aggregate from four of its distinct web archives. To include North Korean texts, data is sourced from Kim Il Sung University[10] (*Journal* and *Literary Works*) and the Korea News Service[11], which archives news from the *Korean Central News Agency* (*KCNA*). The corpus is further supplemented by materials from the Korea Copyright Commission's *GongU Madang*[12], which span a wide historical period, and by modern news articles from five major newspapers (*Kyunghyang Shinmun, Maeil Business Newspaper, The Chosun Ilbo, The Dong-A Ilbo, and The Hankyoreh*) provided by Naver[13].

3.2. Text Preprocessing

Our preprocessing pipeline first applied universal normalization to all documents and then used source-aware heuristics to remove archival and digitization artifacts. The initial normalization stage standardized Unicode text to its NFKC form to resolve compatibility variants, collapsed irregular whitespace, and removed non-printable control characters. The subsequent cleaning stage applied targeted rules to specific sources. These operations included:

- Removing boilerplate text that indicates missing content or metadata.
- Removing modern in-line translations and isolating original texts.
- Stripping erroneous metadata, like modern Korean titles or image captions that were merged into the main text field.
- Filtering out documents with a high ratio of noise-to-text, which often resulted from encoding errors.

[2] https://db.history.go.kr
[3] https://sjw.history.go.kr
[4] https://sillok.history.go.kr
[5] https://www.nl.go.kr
[6] https://archives.hangeul.go.kr
[7] https://db.itkc.or.kr
[8] https://kyudb.snu.ac.kr
[9] https://www.aks.ac.kr
[10] http://www.ryongnamsan.edu.kp
[11] http://www.kcna.co.jp
[12] https://gongu.copyright.or.kr
[13] https://newslibrary.naver.com

Key	Type	Description
id	string	Unique document identifier.
text	string	Main content of the document.
content	object	Raw textual components.
year	integer	Publication year of the document.
language	string	Primary language of the text.
script	string	Primary script used in the text.
source	string	Source institution or archive.
corpus	string	The name of the collection.
copyright	string	Copyright or license status.
url	string	Link to the original document.
format	string	Template for 'text' field.
metadata	object	Object for source-specific fields.
analytics	object	Computed text metrics.

Table 3: JSON Schema. The schema provides fields for identification, classification, text content, and metadata.

3.3. Language and Script Identification

We perform an initial analysis using two methods: a GlotLID (Kargaran et al., 2023) language identification model constrained to predict from a small set of relevant languages (*e.g.*, Korean, Japanese, Chinese), and a character-level script analysis that computes the proportion of Hangul, Hanja, and Kana. The final classification is determined by a set of rules that integrate these preliminary results with source-specific metadata. This heuristic approach is necessary, as modern language identification models are less reliable for historical texts. For instance, documents from sources known to contain only Classical Chinese, such as the *Annals of the Joseon Dynasty*, are categorized as Classical Chinese based on their origin. For Korean-language texts, we assign a historical period—Middle, Early Modern, or Modern Korean—based on the document's publication year.

3.4. Data Schema and Format

The corpus is distributed in the JSON Lines format, where each line is a self-contained JSON object representing a single document. As shown in Table 3, the schema is designed for flexibility and traceability, allowing for the aggregation of documents from diverse sources. Each entry includes fields for identification and attribution, manually curated classification labels for targeted analysis, and distinct fields for both raw and normalized text to ensure data integrity and usability for NLP research. An example of a single data instance in JSON format is shown in Figure 2.

3.5. Corpus Statistics and Distribution

The Open Korean Historical Corpus is a large-scale collection of 17.7 million documents from the 19 sources detailed in Table 2. It contains 6.3 billion characters, which corresponds to 5.1 billion tokens as measured by the `o200k_base` tokenizer[14], making it a substantial resource for the historical study of the Korean language. The corpus's temporal distribution, visualized in Figure 3, shows a heavy concentration in later periods. The apparent data sparsity before 1392 is partly due to visualization constraints, as many individual historical documents, particularly from non-periodical sources like AKS, KLC, and NHMA, lack precise year-level metadata and are therefore not plotted. The data volume increases substantially with the


```
{
  "id": "news_archive:CNTS-00093108108",
  "text": "昨日內部에셔 郡守奏本을 奉呈ᄒᆞ얏더라",
  "content": {
    "body": "昨日內部에셔 郡守奏本을 奉呈ᄒᆞ얏더라",
    "title": "郡奏三度"
  },
  "year": 1905,
  "language": "Modern Korean",
  "script": "Hanja, Old Hangeul",
  "source": "National Library of Korea",
  "corpus": "Korean Newspaper Archive",
  "copyright": "Public Domain",
  "url": "https://www.nl.go.kr/newspaper/detail.do?
content_id=CNTS-00093108108",
  "format": "{body}",
  "metadata": {
    "host_ko": "대한매일신보",
    "year_str": "西曆一千九百五"
  },
  "analytics": {
    "text_length": 19,
    "content_body_length": 19,
    "content_title_length": 4
  }
}
```


Figure 2: An example of a JSON data instance from the Open Korean Historical Corpus.

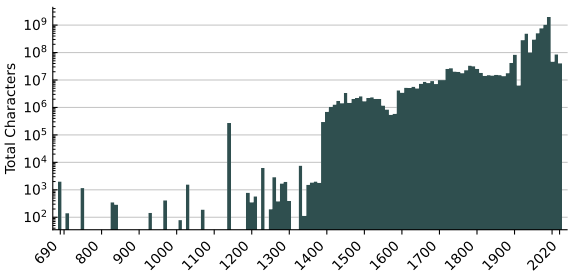

Figure 3: Temporal distribution of the corpus. The total number of characters per decade is plotted on a logarithmic scale. The data volume increases significantly following the establishment of the Joseon dynasty in 1392 and shows a consistent upward trend into the modern era.

[14] https://github.com/openai/tiktoken

establishment of the Joseon Dynasty in 1392, a period known for systematic record-keeping, and then expands exponentially from the late 19th century with the advent of mass media. This results in a sampling bias where modern texts are far more represented than those from earlier eras.

3.6. Legal Considerations

This corpus is distributed under the Creative Commons Attribution-NonCommercial 4.0 International license in compliance with South Korean copyright law, *sui generis* database rights, and the National Security Act. Texts published before 1963, which are in the public domain, are included in full, accounting for 41.9% of all documents and 2.1 billion tokens. For copyrighted materials, such as post-1963 articles or modern translations, we provide the title, metadata, and direct URLs to the original sources to avoid infringement. A detailed breakdown of licensing and distribution formats is provided in Appendix A. The corpus's non-commercial academic and research license explicitly falls within the permissible use exceptions for database rights, allowing reproduction and distribution for educational and scholarly purposes without commercial intent. North Korean texts are also provided as title, URLs, and metadata; their inclusion is strictly for linguistic and scholarly analysis, a non-ideological purpose that complies with the National Security Act.

4. Discussion

4.1. Temporal Dynamics of Korean-Style Sinitic

To quantitatively analyze the temporal distribution of Korean-style Sinitic (Idu), we first compiled a merged lexicon from two Idu dictionaries (DKU, 2025; AKS). Using this lexicon, we identify Idu documents using a longest non-overlapping matching method with the Aho–Corasick algorithm (Aho and Corasick, 1975). Documents are scored by the frequency of Idu markers, normalized by the Hanja character count. A document is classified as Idu if this score exceeds length-stratified thresholds (1.04% for $\leq$100 chars; 0.86% for 101–300 chars; 0.38% for $\geq$301 chars); exclusion lists are employed to filter borderline cases and minimize false positives. To mitigate proportional bias, our analysis is restricted to the 1637–1910 period, which ensures comparable coverage across the four subcorpora (RDGO, DRS, AKS, KIKS) of administrative and legal records; we also exclude sources with an Idu share below 1%.

As shown in Figure 4, the data reveal a gradual increase in Idu usage from 1637 to the 1860s, consistent with higher survival rates for later-period documents (Lee, 2004), followed by a steady decline through 1910. The post-1860s decline indicates factors beyond preservation artifacts. The most significant drop occurs between the 1890s and 1900s, strongly correlating with major language policy changes: the Kabo Reform (1894), which mandated Hangul in official documents, and a 1908 directive favoring Hanja–Hangul script over Idu. This temporal alignment suggests language policy accelerated the decline of Idu in administrative contexts. We interpret these figures as conservative lower bounds, as the corpus includes only digitized materials and our classification method was optimized for high precision.

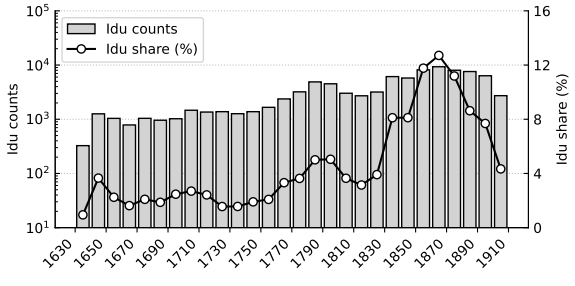

Figure 4: Temporal distribution of Korean-style Sinitic (Idu) usage (1630s-1910s). The bars (left y-axis, logarithmic scale) represent the total count of Idu instances per decade. The black line with circular markers (right y-axis, percentage) illustrates the share of Idu within the analyzed documents from the 1630s to the 1900s.

4.2. The Diachronic Shift from Hanja to Hangul

To measure the diachronic shift of script usage from Hanja to Hangul, we analyzed all documents in our corpus written in the Korean language or

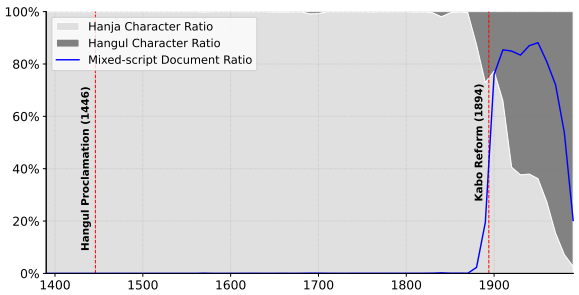

Figure 5: Script transition from Hanja to Hangul between 1390 and 1990, aggregated by decade. The stacked area chart shows the proportion of Hanja (light gray) versus Hangul (dark gray) characters. The blue line indicates the ratio of documents written in Hanja-Hangul mixed script.

Source	OOV Words
NK News	**돐**에, 파**쑈**적인, **뻬**루, 도이**췰**란드, **뷸**레찐은, 도**꾜**, **웥**남사회주의공화국, **줴**쳤다, 브라질주체사상연구**쎈**터, **뽈**스까, **먄**마, 에**꽈**도르, **죤**, **홰**불, **챠**베스, 마**쩨**고라, **낟**알털기를, 바**씰**레, 그**쯘**히, **쟝**, 아제르바이**쟌**공화국, **꼰**스딴찐, 브류**쎌**에서, 어깨**겯**고, **멎**은, **겡**이찌, **꿩**, **윅**또르, **쑬**레이만, 세이**쉘**공화국 anniversary, fascist, Peru, Germany, bulletin, Tokyo, Socialist Republic of Vietnam, shouted, Brazil Juche Idea Study Center, Poland, Myanmar, Ecuador, John, torch, Chavez, Matsegora, grain threshing, Vasile, to that extent, Jean, Republic of Azerbaijan, Constantine, in Brussels, side-by-side, stopped, Genichi, pheasant, Viktor, Suleiman, Republic of Seychelles
SK Web	가리**킵**니다, **옅**은, **챗**봇, **욥**의, **큽**니다, **퀀**텀, 어**댑**터, **퀵**, 몽**쉘**이의, 배**웁**니다, **홋**카이도, **찝**찝한, 느**낍**니다, 왓**챠**, 긍휼을, 보살**핌**을, **멱**살을, **엡**, **훅**, **웇**칠, **헹**궈, 슬**펐**다, 비**젼**을, **켤**레, **쯧**, **갭**, **챕**터, **펭**귄, **쇳**소리, 만**듦** points to, light, chatbot, Job's, is big, quantum, adapter, quick, Mongcael's, learns, Hokkaido, uncomfortable, feels, Watcha, mercy, care, collar, app, hook, lacquer, rinse, was sad, vision, pair, tsk, gap, chapter, penguin, metallic sound, making

Table 4: Top 30 most frequent OOV word stems from the NK News and SK Web corpora, as identified by the KLUE-BERT model. The table displays the most frequent example word that corresponds to each stem (in bold), followed by its English translation.

Model	Type	NK News	SK Web
mBERT	M	0.2823	0.3196
KLUE-BERT	K	0.2035	0.0650
KcBERT	K	0.1226	0.0024
XLM-R	M	0.0499	0.0125

Table 5: Out-of-Vocabulary rates for North Korean (NK) and South Korean (SK) text. The values represent the percentage of unknown tokens generated by four different encoder models when tokenizing the NK News and SK Web corpora. Type indicates if the model is trained on multilingual (M) or South Korean (K) texts.

Classical Chinese from 1390 to 1999. Using regular expressions based on Unicode properties and blocks, we counted Hanja and Hangul characters in each document. These counts were then aggregated into a timeframe of 10-year decades, and we calculated both the character-level script proportion and the ratio of documents written in the Hanja-Hangul mixed script system, where the threshold is over 10% of both Hanja and Hangul usage.

As shown in Figure 5, the data reveals a rapid transformation rather than a gradual shift. From the release of Hangul in the mid-15th century to the late 19th century, Hanja usage was dominant, with almost all documents written exclusively in Classical Chinese. A sharp decline in Hanja-exclusive texts began around 1890, which can be interpreted as a result of modernizing efforts such as the Kabo Reform (1894), which mandated Hangul for official documents. Meanwhile, as Hanja-Hangul mixed script usage rose and then fell, the usage of Hangul increased rapidly. By the 1980s, Hangul became the primary script, accounting for over 93% of characters and reflecting language policies that relegated Hanja to a supplementary role. This result accurately reflects known turning points of Korean language history, and shows that our corpus is useful for such diachronic analysis.

4.3. A Preliminary Analysis of Tokenizer Coverage on North Korean Text

Decades of political separation have caused the Korean language to diverge lexically between the North and South. As a preliminary step to understand potential model limitations, we assess how well the tokenizer, the initial stage of language modeling, handles this divergence. We quantify this gap by measuring the out-of-vocabulary (OOV) rate—the proportion of unknown tokens generated by the tokenizers of four pre-trained encoder models[15]—when processing our North Korean news corpus (KCNA) and a comparably-sized South Korean web text baseline.[16] All texts were pre-processed using NFKC normalization and filtered to retain only complete Hangul syllables.

Our results confirm a measurable lexical gap at the token level. As shown in Table 5, OOV rates for North Korean text were significantly higher for most models, ranging from 3 to 51 times, though the mBERT model showed a comparable rate. However, the absolute rates remained low (under 0.3%), suggesting modern subword tokenizers are surprisingly resilient in segmenting the text.

[15]Model identifiers are `google-bert/bert-base-multilingual-cased`, `FacebookAI/xlm-roberta-base`, `klue/bert-base`, and `beomi/kcbert-base`.

[16]The South Korean baseline was a 5% random sample of the `HAERAE-HUB/KOREAN-WEBTEXT` corpus, which is available on Hugging Face. Our KCNA corpus contains approximately 170k documents and 126M characters; the sampled SK corpus contains approximately 64k documents and 175M characters.

An analysis of the most frequent OOV tokens (Table 4) reveals the divergence stems primarily from North Korea's unique orthography for loanwords (*e.g.*, 도이췰란드 for Germany) and distinct vocabulary (*e.g.*, 돐 for anniversary). This lexical gap, though small, can still disproportionately impact sensitive tasks like named entity recognition, highlighting a clear opportunity for model improvement through vocabulary expansion.

Several limitations apply to this analysis. The two corpora differ in domain: North Korean news articles versus South Korean general web text, which means topic and genre differences may inflate OOV rates beyond what is attributable to linguistic divergence alone. Furthermore, North Korean state media is narrow in both topical coverage and stylistic diversity, and may not represent everyday North Korean language use. The reported OOV rates should therefore be interpreted as an upper bound reflecting both true lexical divergence and domain mismatch.

5. Related Work

Korean Corpora for Modern NLP Large-scale multilingual corpora have incorporated Korean data, but their coverage remains limited in both temporal and linguistic diversity, failing to capture earlier stages or varieties of the language that are essential for studying linguistic change and variation. Resources such as OSCAR (Ortiz Suárez et al., 2019), mC4 (Xue et al., 2021), and CulturaX (Nguyen et al., 2024) include Korean data, but these materials primarily reflect modern standardized usage. Dedicated Korean corpora, including AI-Hub[17] (NIA, 2020), koTenTen (Sketch Engine, 2018), Korpora (ko-nlp team, 2019), and the Open Korean Corpora (Cho et al., 2020), have established valuable resources for modern Korean NLP. However, their scope likewise remains confined to contemporary language. Moreover, access to AI-Hub datasets is limited to domestic applicants, and their overseas use—including by Korean nationals abroad—is either restricted or requires separate agreements with the responsible institutions and government bodies, limiting global research use. Overall, existing Korean corpora lack temporal depth and script variation, leaving historical materials and mixed-script texts largely unrepresented.

Historical Korean Corpora Historical corpora play a crucial role in preserving cultural heritage and enabling longitudinal analyses of language and society. Historical corpora initiatives such as Shamela (Belinkov et al., 2016), CHisIEC (Tang et al., 2024), and CCOHA (Alatrash et al., 2020) illustrate how historical datasets can mitigate temporal bias and support research on long-term linguistic and cultural change.

In the Korean context, the inclusion of classical texts written in Hanja has long been regarded as essential, since Hanja served as the principal written medium prior to the invention and popularization of Hangul. Several efforts have been made to build historical corpora, including the Integrated Database of Korean Classics (ITKC) and KoHiCo (NIKL, 2024). The lack of standardized formats and limited script coverage, however, still pose challenges for researchers and hinder reproducibility.

National projects further contribute by encompassing both classical and modern data, but their main challenge lies in restricted accessibility and licensing. The Sejong Corpus (1998–2007) consists of about 200M tokens across written, spoken, parallel (Korean–English/Japanese), and historical data (An et al., 2025). However, some obstacles, such as its limited distribution (mainly in DVD format), have hindered its usage. The Modu Corpus (NIKL, 2019) provides diverse datasets (sentiment, NER, parallel texts, newspapers, everyday dialogues) and serves as a de facto standard resource. However, access requires registration and approval, and its diachronic coverage remains narrow.

To overcome limited accessibility, openness, and script coverage of the existing historical corpora, we introduce an openly available corpus spanning a broad temporal and linguistic range —including *Hanja*, Korean-style Sinitic (*Idu*), Middle and Early Modern Korean, and Japanese texts from the colonial period—integrated into a unified, NLP-ready format for large-scale exploration of Korean language and culture across time.

6. Conclusion

In this work, we introduced the Open Korean Historical Corpus, a large-scale, openly licensed dataset of 17.7 million documents designed to address the critical lack of accessible historical data for Korean NLP. By compiling and standardizing texts from the 7th century to the present, our corpus provides the first comprehensive resource for studying the language's diachronic evolution across diverse languages and writing systems, including Classical Chinese, Korean-style Sinitic, and Hanja-Hangul mixed script. We present quantitative analyses of notable linguistic changes, including the temporal dynamics of Korean-style Sinitic, the Hanja-to-Hangul transition, and the lexical divergence of North Korean.

Beyond these analyses, the corpus opens up

[17] https://www.aihub.or.kr

practical possibilities in an area where current Korean NLP tools fall short: they are predominantly trained on modern Hangul and perform poorly on historical texts. Our resource could serve as a foundation for closing this disconnect — for instance, by training tokenizers and language models that are robust to diachronic variation in script and orthography, pre-training specialized encoder models for downstream digital humanities tasks on archival materials, or continually pre-training large language models to improve their handling of historical writing systems. More broadly, the corpus's temporal depth can support historical linguistics research by enabling fine-grained diachronic analyses across genres, regions, and document types. We release the corpus and our processing code to facilitate such research and to lay the groundwork for making centuries of historical documents more accessible.

Acknowledgements

This research was conducted as part of the Sovereign AI Foundation Model Project(GPU Track), organized by the Ministry of Science and ICT(MSIT) and supported by the National IT Industry Promotion Agency(NIPA), S.Korea. (PJT-26-010017)

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the Top-Tier AI Global HRD invitation program (RS-2025-25461932) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research.

Ethics Statement

The Open Korean Historical Corpus is provided for research in historical linguistics and natural language processing. To ensure academic integrity, we provide detailed source information for all texts to properly attribute the work of the original creators and digitizing institutions.

We caution users to handle the data with care, as it contains sensitive materials rooted in outdated cultural norms, including gender, religious, class, and regional discrimination. We have included North Korean texts from state media for linguistic analysis only; this inclusion does not constitute an endorsement of their ideological content, which features political propaganda, censorship, extreme language, and a cult of personality. We also acknowledge that historical documents contain biases that language models trained on this corpus may learn and reproduce. To prevent the downstream harms of naive or blind usage, researchers must actively consider and mitigate these biases during model training.

Limitations

The primary limitation of this corpus is its inherent sampling bias. It is constructed from extant and digitized materials, which do not represent a complete or balanced sample of all historical documents ever produced. This bias is evident in the temporal distribution of the data, where modern periods are heavily overrepresented compared to earlier eras from which fewer texts have survived or been digitized. Furthermore, the scope of the corpus is intentionally limited to written texts. It does not include transcribed spoken language, making it unsuitable for certain linguistic studies, such as those focusing on phonological or conversational phenomena.

Bibliographical References

Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340.

AKS. 한국학중앙연구원 이두용례사전 [AKS Idu Example Dictionary]. Accessed on 2025-10-14.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Jinsan An, Kilim Nam, Hyeon ah Kang, and Jun Lee. 2025. 말뭉치언어학의 관점에서 본 〈모두의 말뭉치〉 구축 현황과 쟁점 [Everyone's Corpus: Corpus Linguistics Issues]. 한글 [*Han-geul*], 86(2):489–525.

Olivier Bailblé. 2016. History of the dative markers in Korean language: From old Korean to contemporary Korean. *International Journal of Korean Humanities and Social Sciences*, 1:119–136.

Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. 2016. Shamela: A large-scale historical Arabic corpus. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities* (*LT4DH*), pages 45–53, Osaka,

Japan. The COLING 2016 Organizing Committee.

Won Ik Cho, Sangwhan Moon, and Youngsook Song. 2020. Open Korean corpora: A practical report. In *Proceedings of Second Workshop for NLP Open Source Software* (*NLP-OSS*), pages 85–93, Online. Association for Computational Linguistics.

Sungjae Choo. 2016. The use of Hanja (Chinese characters) in Korean toponyms: Practices and issues. *Onoma*, 51:13–24.

Robert Crellin and Terry Joyce. 2019. Writing systems: Past, present (··· and future?). *Written Language & Literacy*, 22(2):167–178.

DKU. 2025. 단국대학교 이두사전 [Dankook University Idu Dictionary]. Accessed on 2025-10-14.

Ik-sang Eom. 2002. The origin of Sino-Korean coda. *Korean Linguistics*, 11(1):101–117.

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

Yong-joo Hwang and Jeong-do Choi. 2016. 21 세기 세종 말뭉치 제대로 살펴보기-언어정보나눔터 활용하기 [Looking properly at the 21st Century Sejong Corpus: Utilizing the language information sharing center]. 새국어생활 [*New Korean Life*], 26(2):73–86.

ITKC. 한국고전종합db [Integrated Database of Korean Classics]. Accessed on 2025-09-02.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Dongso Kim. 1998. 한국어 변천사 [*The History of Korean Language Variation*]. Hyungseul Publishing. (in Korean).

Moo-rim Kim. 2004. 국어의 역사 [*History of the Korean Language*]. Hankookmunhwasa, Seoul. (in Korean).

Ross King. 2022. Idu in and as Korean literature. In Heekyoung Cho, editor, *The Routledge Companion to Korean Literature*, pages 123–140. Routledge.

ko-nlp team. 2019. Korpora: Korean corpora archives. Accessed: 2025-09-02.

Ki-Moon Lee and S Robert Ramsey. 2011. *A history of the Korean language*. Cambridge University Press.

Kwankyu Lee. 2021. 남북한 어문 규범의 변천과 과제 [*The Changes and Tasks of Language Norms in South and North Korea*]. Korea University Press.

Sugeon Lee. 2004. *16*세기 한국 고문서 연구 [*Studies of Korean old documents in the sixteenth century*]. Akanet, Seoul. (in Korean).

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*), pages 4226–4237, Torino, Italia. ELRA and ICCL.

NIA. 2020. AI-Hub: Artificial intelligence data service. (in Korean).

NIKL. 2010. 우리말의 이모저모 [*Everything You Wanted to Know about the Korean Language*]. Human Culture Arirang, Seoul. (in Korean).

NIKL. 2019. Modu corpus. (in Korean).

NIKL. 2024. Korean historical corpus (KoHiCo). Accessed: 2025-09-02.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora* (*CMLC-7*), pages 9–16. Leibniz-Institut für Deutsche Sprache.

Hye K. Pae, editor. 2018. *Writing Systems, Reading Processes, and Cross-Linguistic Influences*. John Benjamins Publishing Company.

Soon-Yong Pak and Keumjoong Hwang. 2011. Assimilation and segregation of imperial subjects: “educating” the colonised during the 1910–1945 Japanese colonial rule of Korea. *Paedagogica Historica*, 47(3):377–397.

Sketch Engine. 2018. koTenTen: Corpus of the Korean web. Accessed: 2025-09-02.

Jae Jung Song. 2015. Language policies in North and South Korea. In Lucien Brown and Jaehoon Yeon, editors, *The Handbook of Korean Linguistics*, pages 477–491. Wiley, Hoboken, NJ.

Xuemei Tang, Qi Su, Jun Wang, and Zekun Deng. 2024. CHisIEC: An information extraction corpus for Ancient Chinese history. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*), pages 3192–3202, Torino, Italia. ELRA and ICCL.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Appendix

A. Data Availability and Licensing

Source	License				Distribution	
	All Rights Reserved	Public Domain	CC BY-NC-ND 2.0 KR	KOGL Type 1	Text & Metadata	Metadata Only
NNL	10,063,510	3,472,984	0	0	25.7%	74.3%
DRS	0	1,792,187	0	0	100.0%	0.0%
KLC	0	652,405	0	0	100.0%	0.0%
KCNA	170,472	0	0	0	0.0%	100.0%
KNA	0	364,409	0	0	100.0%	0.0%
AJD	0	413,131	0	0	100.0%	0.0%
RDR	0	338,084	0	0	100.0%	0.0%
RDGO	0	130,143	0	0	100.0%	0.0%
MCM	0	15,326	0	0	100.0%	0.0%
GUM	0	13,291	0	0	100.0%	0.0%
JKU	39,723	0	0	0	0.0%	100.0%
AKS	0	55,409	73	0	100.0%	0.0%
RBDC	0	93,528	0	0	100.0%	0.0%
RJL	0	22,502	0	0	100.0%	0.0%
KIKS	0	32,487	0	0	100.0%	0.0%
HG	0	20,047	0	0	100.0%	0.0%
NHMA	0	1,654	0	15	100.0%	0.0%
LKU	274	0	0	0	0.0%	100.0%
HTK	0	4,613	0	0	100.0%	0.0%
Total	**10,273,979**	**7,422,200**	**73**	**15**	**41.9%**	**58.1%**

Table 6: Legal licensing and distribution formats across the 19 sources in the Open Korean Historical Corpus. Documents are categorized into four license types, which dictate their available distribution format. Full text is provided for openly licensed materials (Public Domain, CC BY-NC-ND 2.0 KR, KOGL Type 1), while copyrighted sources (All Rights Reserved) are restricted to metadata and URLs.

Table 6 presents the licensing and distribution format for all 19 sources. Overall, 41.9% of documents have full text included, while 58.1% are metadata-only, containing title, year, language, and a URL to the original source. The metadata-only category consists of copyrighted ("All Rights Reserved") materials for which we lack redistribution rights. The vast majority of these—10.1 million out of 10.3 million—come from the Naver News Library (post-1963 newspaper articles). The remaining copyrighted entries are the three North Korean sources (KCNA, JKU, LKU), for which metadata-only distribution also ensures compliance with the National Security Act.

Among openly licensed materials, most documents are in the public domain. Two additional license types appear in small quantities: 73 documents from AKS under CC BY-NC-ND 2.0 KR, a non-commercial, no-derivatives license; and 15 documents from NHMA under KOGL (Korea Open Government License) Type 1[18], an attribution-only public license issued by the South Korean government that is functionally equivalent to CC BY but not officially declared interoperable with Creative Commons.

[18] https://www.kogl.or.kr/info/license.do