

Cellik Adams,¹ EunKyoung Jo,^{1*} Juae Kim²

²Hankuk University of Foreign Studies, Republic of Korea

cadams@sogang.ac.kr, eunyoung@sogang.ac.kr, juaekim@hufs.ac.kr

Language models learn semantic similarity from shared context; however, when a single word has multiple orthographies that rarely appear in the same context, models cannot learn their equivalence. This is a structural problem that cannot be solved with scale. Korean presents a clear case: 20th-century print media used Hanja (Sino-Korean script), while contemporary models train on Hangul-only (Korean alphabet) text. Multilingual models inherit script segregation from their training distribution, where Hanja appears almost exclusively outside Korean text, and monolingual Korean models assign separate vocabulary entries to script variants, forcing distinct representations from initialization. We introduce Han2Han, a 169M parameter encoder-decoder model trained with a language-aware recipe—jamo and character-level embedding fusion, morpheme-aware denoising, and bidirectional Hanja-Hangul transcription—that forces orthographic alignment as a prior to distributional learning. The resulting representations are script-invariant: the same content receives the same representation regardless of script. On temporal classification of mixed-script historical newspapers, evaluated on an article-level held-out split, Han2Han reaches 76.0 macro F1 against 54.3 for the strongest baselines (KLUE-BERT and the $4.7\times$ larger T5Gemma 2 alike), a 21.7-point margin, demonstrating that language-specific pre-training yields capabilities not addressed by current Korean-inclusive multilingual training at this scale.

Language models learn semantic similarity from shared context, but this creates a structural limitation: when the same word can be written in multiple orthographies that rarely co-occur in training data, models have no mechanism to learn their

[illegible]

equivalence. A clear case is offered with the Korean language, whose print media for most of the 20th century represented Sino-Korean vocabulary with Hanja for explicit semantic disambiguation. In contrast, contemporary Korean is Hangul-only and Hangul is the orthography on which most language models are trained. Multilingual models inherit script segregation from their training distribution; monolingual Korean models, trained on Hangul-dominant corpora, encounter Hanja too rarely to learn script equivalence. More data only adds more examples of each variant in its distinct context, which reinforces rather than bridges this separation.

Han2Han addresses this through architectural intervention. Jamo-level embeddings exploit the systematic correspondence between Hanja and Hangul by forcing orthographic alignment through shared subword features before distributional learning begins. Morpheme-aware masking respects Korean’s agglutinative morphology, creating corruption spans aligned with linguistic boundaries rather than arbitrary token sequences. Bidirectional

Hanja-Hangul transcription explicitly teaches the model that 美術 and 미술 represent identical semantic content. These Korean-specific techniques encode linguistic knowledge directly into the architecture; the principle generalizes to other languages with structure that can be encoded *a priori* rather than discovered through training, enabling capabilities unattainable through scale alone.

This work makes three primary contributions. First, we identify the training recipe—jamo and character-level embedding fusion, morpheme-aware denoising, and bidirectional Hanja-Hangul transcription—that yields script-invariant Korean representations. Off-the-shelf models trained at substantially larger scale (T5Gemma 2 at 786M parameters, trained on a 2T-token Korean-inclusive UL2 mixture) segregate documents entirely by script (Figure 3, Table 9); applying the Han2Han objective via continued pre-training produces improved script-invariance in T5Gemma 2, showing that the recipe is portable to existing pre-trained models. Second, we deliver this capability at 169M parameters trained from scratch with comparable KLUE benchmark performance at 35B tokens seen versus 131B–6.3T tokens for baselines, and a 21.7-point macro-F1 margin over the strongest baselines on temporal classification of mixed-script historical newspapers. Third, we enable computational analysis of mixed-script Korean texts, providing digital humanities researchers access to centuries of Korean intellectual production without the need for institutional ML infrastructure. Code, model weights, and evaluation splits are available at <https://github.com/cadazar/han2han>.

2 Background

2.1 Sino-Korean Characters (Hanja) and the Korean Characters (Hangul)

The Korean writing system, *hangul* (한글), is an alpha-syllabic script that combines features of both alphabetic and syllabic writing systems. Each Hangul syllable block is composed of individual letters called *jamo* (자모), which can represent either consonants or vowels. Korean before the 2000s frequently includes *hanja* (한자)—Sino-Korean logographic characters used to represent Sino-Korean words—which compose an estimated 60-70% of Korean vocabulary. Each Hanja character maps to a single Hangul syllable, but due to the finite phoneme inventory, a single Hangul syllable may correspond to dozens of distinct Hanja characters

with different meanings (e.g., 수 can represent 水 "water," 手 "hand," 數 "number," among others). The relationship also runs in the opposite direction: the same Hanja character may take different Hangul readings depending on context, as in 金 read 금 in 金融 (*finance*) but 김 in 金浦 (*Gimpo*), and 女 read 녀 in 婦女 (*woman*) but 여 in 女子 (*woman*).

Korean’s mixed-script history is one instance of digraphia. Others are synchronic—Serbian in Cyrillic and Latin, Hindi-Urdu across Devanagari and Perso-Arabic—or diachronic script reforms, of which the Ottoman-to-Republican Turkish transition, a literate tradition rendered unreadable within a few decades, is the closest parallel to Korea’s. Simplified and traditional Chinese are a third case, one language in two character inventories. What separates Korean is that the mapping is not recoverable by rule in either direction, as shown by the examples above. Serbian and simplified-traditional Chinese admit near-deterministic transliteration, so a normalization pass can bridge the variants before a model sees them; for Korean no such pass exists and the equivalence must be learned for unambiguous semantic representation. This is why we treat orthographic alignment as a pre-training objective rather than a preprocessing step.

While contemporary language models disambiguate homophones like 수 effectively using local context, mixed-script Korean poses a more fundamental challenge that context alone does not address. Because tokenizers assign distinct vocabulary entries to semantically equivalent terms differing only in script (somewhat similar to treating "color" and "colour" as unrelated words), the same word occupies disjoint regions of the embedding space depending on how it happens to be written. For document-level representations of monolingual Korean corpora that mix scripts, this means documents are organized by orthography rather than meaning: semantically similar documents written in different scripts receive dissimilar representations. We demonstrate this effect directly in Figure 3 (Section 5.2), and the rest of this paper describes Han2Han’s response: to learn script-invariant representations that capture the etymological relationships implicit in Korean’s writing system evolution.

2.2 Related Work

Prior Korean NLP work that engages with the writing system clusters into two largely disjoint

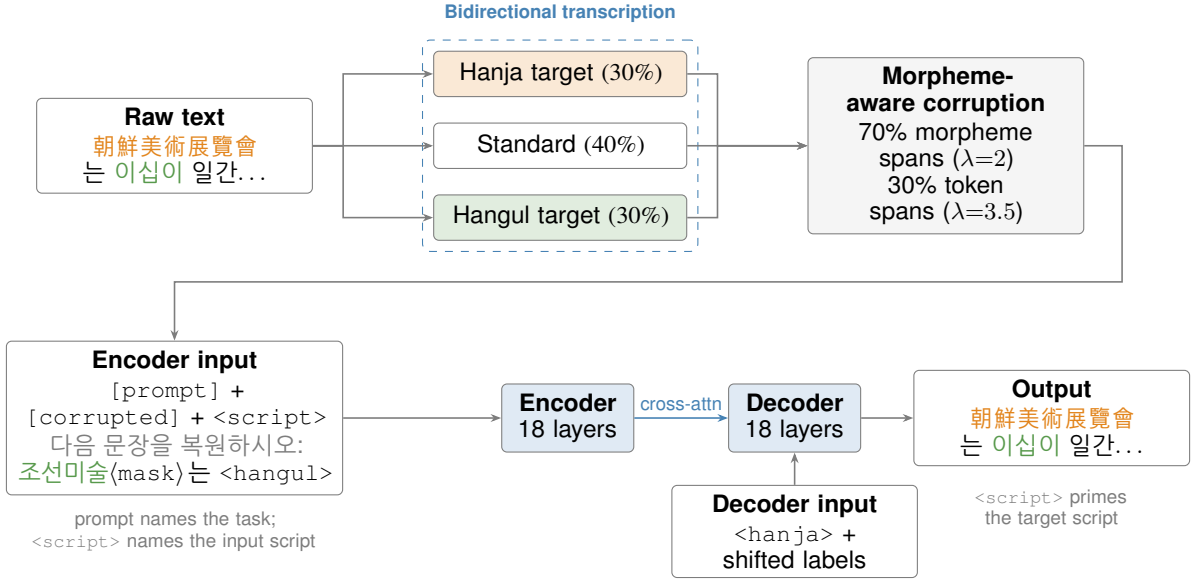


Figure 2: **Han2Han pre-training scheme.** Each training example is routed through one of three transcription modes—**Hanja target**, where the decoder predicts the original Hanja; **Standard** denoising, where scripts are unchanged; and **Hangul target**, where labels are transcribed to Hangul—so that the same content is seen in both writing systems across the corpus. Corruption then masks spans at McCab-Ko morpheme boundaries (70%) or at arbitrary token boundaries (30%). A natural-language task prompt prepended to the encoder input conditions the model on the denoising mode, replacing UL2’s opaque paradigm-indicator tokens (Section 3), and script tokens (<hanja>, <hangul>) on the encoder and decoder inputs name the input and target writing systems. The design teaches the model that 朝鮮美術展覽會 and 조선미술전람회 carry identical semantic content in different scripts.

strands: Hanja-aware modeling targeted at pre-modern or cross-lingual settings, and subcharacter tokenization for contemporary Hangul-only text. Each strand contributes part of the apparatus Han2Han needs, but neither addresses the modern mixed-script Korean that motivates this work—text roughly spanning the 20th century in which Hanja and Hangul co-occur within the same document and frequently within the same sentence.

In the first strand, Yoo et al. (2019) augment Korean word embeddings with Hanja signal at the type level, showing that Hanja carries semantic information not recoverable from Hangul alone. Yoo et al. (2022) and Son et al. (2022) build pre-trained models, benchmarks, and translation systems over premodern Hanja documents—chiefly the Annals of the Joseon Dynasty—and Yang et al. (2023) release a Joseon-era document-level relation extraction dataset. Kim et al. (2020) exploit Hanja-Kanji vocabulary overlap to improve Korean-to-Japanese neural machine translation. These works treat Hanja as an enrichment signal for Hangul word vectors, as the dominant script of a closed premodern corpus, or as a cross-lingual bridge to another Sinographic language; none address the within-document mixed-script setting in which the same

content alternates between Hanja and its Hangul transcription.

In the second strand, work on Hangul subcharacter representation decomposes syllables into sub-syllabic units to share parameters across morphologically related forms. Park et al. (2018) adapt fastText (Bojanowski et al., 2017) to Hangul jamo n -grams for static word vectors, and Seonwoo et al. (2019) extend the same line of work. Cognetta et al. (2023) build a parameter-efficient character-level Korean language model exploiting Hangul’s compositional structure; Jeon et al. (2023) combine linguistically informed subword tokenization with sub-character decomposition for masked language modeling; Li and Zhang (2025) pursue the analogous decomposition for Chinese, supervising character recognition jointly against ideographic description sequences of radicals; the sub-character unit there plays much the same role our jamo and character n -grams do, though the task (recognition from images) and the direction of the decomposition differ. Kim et al. (2024b) encode the combination rules of subcharacters explicitly in the vocabulary; and Lee et al. (2025) show that jamo-level tokenization outperforms standard tokenization in low-resource Korean translation. These approaches

collectively establish that sub-syllabic structure is exploitable for Korean, but they operate within a single script: a Hanja-written word and its Hangul transcription share no subword units, so the orthographic gap between scripts is not bridged at the subword level.

Han2Han is, to our knowledge, the first public-facing Korean language model to combine both strands within a single pre-training recipe. We perform Hanja→Hangul transcription before subword decomposition, so that Hanja and Hangul cognates such as 韓國語 and 한국어 project to identical jamo n -gram features, and we pair this with a bidirectional Hanja↔Hangul denoising objective that turns the resulting embedding alignment into a learning signal during pre-training. Previous Korean NLP efforts have targeted either contemporary Hangul-only text (Park et al., 2021) or premodern Hanmun (Yang et al., 2023); Han2Han is designed for the modern mixed-script period that sits between them. We evaluate on the Korean Language Understanding Evaluation (KLUE) benchmark (Park et al., 2021), building on foundational work in Korean NLI and STS (Ham et al., 2020), and on a temporal classification task constructed from mid-twentieth-century mixed-script newspaper articles (Section 5.3).

2.3 Encoder-Decoder Models

Han2Han adopts the encoder-decoder architecture because efficient natural language understanding, rather than open-ended generation, is the primary objective. UL2 (Tay et al., 2022) showed that encoder-decoder models consistently outperform decoder-only counterparts on NLU benchmarks at matched compute, and T5Gemma (Zhang et al., 2025a) reports substantial gains when adapting decoder-only models into encoder-decoder format. Earlier denoising work (Raffel et al., 2020; Lewis et al., 2020) left open whether their span-corruption recipes were optimal for NLU; our hyperparameter sweeps (Appendix E) find that language-specific denoising (morpheme-aware masking, bidirectional transcription) outperforms generic approaches.

3 Training Data and Objective

3.1 Sources

Training data combines (1) a collected newspaper corpus (1.74B tokens; 1920s–1990s, *Chosun Ilbo*, *Dong-a Ilbo*, *JoongAng*); (2) specialized-domain

Korean corpora (patents, legal documents, court cases, academic papers, bills) from the Korean National Institute of Korean Language; (3) the Korean subset of FinePDFs; (4) the OKHC historical Korean corpus; (5) the Modu Corpus (모두의 말뭉치) spanning written, spoken, news, and messenger text; (6) curated web/encyclopedia sources (OSCAR-Korean, Korean Wikipedia, NamuWiki); (7) the National Library of Korea book corpus across 74 KDC categories; and (8) Blue House Korean Petitions. Together these sources span roughly a millennium of recorded Korean writing through 2025. Full corpus statistics are provided in Appendix A.

3.2 Training Objective

We train with a mixture-of-denoisers following UL2 (Tay et al., 2022), comprising (1) T5-style span masking with sentinel tokens (Raffel et al., 2020); (2) morpheme-aware span masking respecting Korean morphological boundaries via MeCab-Ko with Hanja-aware preprocessing (Appendix B), inspired by TSDAE’s word-deletion approach (Wang et al., 2021) but operating on Korean morpheme spans; (3) spacing¹ and sentence-boundary reconstruction where morpheme-masked inputs contain neither spaces nor sentence delimiters, forcing the model to learn Korean word and sentence boundaries that are notoriously inconsistent in both historical and contemporary text; (4) bidirectional Hanja↔Hangul transcription applied across the R- and X-denoiser variants; and (5) byte-to-token reconstruction following ByT5 (Xue et al., 2022), in which the encoder receives a fully byte-converted sequence and the decoder reconstructs the original tokenization, providing universal script coverage and an additional level of text abstraction.

Concretely, we apply UL2’s R-, X-, and S-denoisers at a 2:2:1 mixing ratio. Following Zhang et al. (2025b), we specify each denoiser by (μ, r, n) —mean span length, corruption rate, and number of corrupted spans (derived from μ, r , and sequence length L). R-denoiser samples uniformly from $\{(3, 0.15, n), (8, 0.15, n), (3, 0.30, n)\}$; X-denoiser samples uniformly from $\{(3, 0.5, n), (8, 0.5, n), (64, 0.15, n), (64, 0.5, n)\}$; S-denoiser is prefix-LM with a single masked continuation, $\mu \sim \text{Uniform}(0.6L, 0.8L)$, $r = \mu/L$, $n = 1$. Within R-denoiser, 70% of sequences replace token-level Poisson spans

¹"띄어쓰기" in Korean.

with morpheme-aware spans of mean morpheme length $\mu_{\text{morph}} \in \{2, 3, 2\}$ (matched by index to the same three r values). All mode ratios, (μ, r) pairs, the morpheme-variant rate, and the Hanja $\leftrightarrow$ Hangul transcription rate within R- and X-denoiser were jointly ablated (Appendix E).

Following UL2’s paradigm-token approach (Tay et al., 2022), we condition the model on the given denoising task; however, we replace the UL2 paradigm indicator tokens with natural-language task prompts prepended to every sequence, which the model parses through its embedding pipeline. These prompts function as richer-signal BOS markers carrying linguistic content rather than as opaque per-task indicators. We find that this conditioning creates a measurable influence on training: a run with inadvertently non-Korean prompts began to diverge after several billion tokens of Korean-only training data; switching to Korean-only prompts resolved the divergence immediately. The signal the model learns to depend on is not merely the presence of a task marker but the linguistic content of the prompt itself. We omit prompts in a final 20% cooldown phase (Gao et al., 2025) so the model functions at inference without explicit task conditioning.

For examples that carry year metadata (approximately 5% of training sequences, predominantly newspaper articles), we replace the S-denoiser with a temporal continuation variant: the encoder is prompted with a period-aware instruction² followed by the prefix, and the decoder produces both a period-appropriate continuation and the year of composition. This couples linguistic style with temporal grounding directly in the pre-training objective.³

3.3 Supervised Fine-Tuning

Consistent with prior findings that instruction tuning broadly improves downstream task performance (Chung et al., 2024; Longpre et al., 2023), we apply a supervised fine-tuning stage to the pre-trained checkpoint prior to all downstream evaluations. Han2Han is fine-tuned on a mixture of

²Korean: “작성 시대를 고려하여 텍스트를 이어 쓰고, 작성 연도를 추정하시오.”. English gloss: “Considering the period of composition, continue the text and estimate the year of composition.”

³The temporal-classification evaluation set (Section 5.3) is an article-level held-out split, registered in a URL exclusion list that the pre-training data loader drops before any training batch is drawn. That mechanism post-dates the released checkpoint; Section 5.3 states the residual exposure this leaves and bounds its effect.

supervised tasks including KLUE benchmark tasks, summarization, and Hanja $\leftrightarrow$ Hangul transcription, with each task paired with natural-language instructions drawn from a small hand-authored pool of 4–5 paraphrases. The fine-tuning objective combines confidence-aware scheduled sampling (Liu et al., 2021) with a token-level contrastive loss between the gold and scheduled-sampling forward passes (He et al., 2024), using their default hyperparameters; regularization follows FLAN’s recommendations (dropout 0.05 throughout, except with additional label smoothing 0.05) with a learning rate of 1×10^{-5} .

4 Architecture

Han2Han is a 169M-parameter transformer encoder-decoder. Depth and dimensions follow Gemma 3’s smallest configuration (Gemma Team, 2025), and the overall encoder-decoder structure is closest to T5Gemma 2 (Zhang et al., 2025b)—specifically its 270m-270m variant. Two choices distinguish Han2Han: encoder and decoder share their main parameters (attention and FFN kernels, token embeddings, and the language-modeling head), and the decoder retains separate cross-attention rather than T5Gemma 2’s merged-attention design. The subtoken (jamo and character) embeddings introduced in the next subsection are fused into the input representation but kept as parameters separate from the shared token matrix to preserve expressivity under tying. Full architecture, attention pattern, optimizer, and training hyperparameters are deferred to Appendix C.

4.1 Sub-token Embeddings

Decomposed Hangul Syllable Embeddings

Subword n -gram embeddings (Bojanowski et al., 2017) represent each word w as the sum of vector embeddings of its character n -grams, $\mathbf{w} = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g$, with $n \in \{3, \dots, 6\}$ and $K = 2 \times 10^6$ hashed embedding buckets in the original formulation. Han2Han adapts this scheme to Hangul’s compositional structure with two principled departures. First, we operate on *jamo* (sub-syllabic phonetic atoms) rather than full characters, extending the original character-level decomposition to subcharacter granularity: each Hangul syllable decomposes into onset, nucleus, and coda, with a hyphen marking empty codas (Figure 4). Second, we use $n = 3$ only, rather than the original slid-

ing range.⁴ This choice is linguistically motivated: each Hangul syllable comprises 2–3 jamo (onset, nucleus, optional coda), normalized to 3 symbols by the empty-coda placeholder –, so every 3-gram captures one syllable—the natural morphophonological unit for Korean. We additionally omit the original word-boundary symbols (</>) because our Unigram tokenizer already prefixes start-of-token markers. With $K = 7K$ jamo hash buckets—two orders of magnitude smaller than a literal port of the original scheme—the jamo embedding matrix $\mathbf{W}_{\text{jamo}}$ stays compact while preserving the parameter-sharing benefit that Bojanowski et al. (2017) originally identified for morphologically related word forms.

Single-Syllable and Token Embedding Fusion

Many Korean morphemes are composed of single-syllable units, particularly Sino-Korean characters, so we add character-level embeddings alongside the jamo n -grams; the three embedding sources are fused with the token embedding as:

Given a token t with jamo 3-gram indices $\{j_1, \dots, j_k\}$ and character indices $\{c_1, \dots, c_m\}$:

$$\mathbf{e}_{\text{tok}} = \mathbf{W}_{\text{tok}}[t] \in \mathbb{R}^d \quad (1)$$

$$\mathbf{e}_{\text{jamo}} = \sum_{i=1}^k \mathbf{W}_{\text{jamo}}[j_i] \in \mathbb{R}^{d/2} \quad (2)$$

$$\mathbf{e}_{\text{char}} = \sum_{i=1}^m \mathbf{W}_{\text{char}}[c_i] \in \mathbb{R}^{d/2} \quad (3)$$

$$\mathbf{f} = \mathbf{e}_{\text{jamo}} + \mathbf{e}_{\text{char}} \quad (\text{subword features}) \quad (4)$$

$$\mathbf{g} = \text{SiLU}(\mathbf{f}) \odot \mathbf{f} \quad (\text{self-gating}) \quad (5)$$

$$\mathbf{h} = \mathbf{e}_{\text{tok}} + \mathbf{W}_{\text{proj}}\mathbf{g} \quad (\text{residual fusion}) \quad (6)$$

$$\mathbf{h}_{\text{out}} = \text{RMSNorm}(\mathbf{h}) \quad (\text{output embedding}) \quad (7)$$

where $\mathbf{W}_{\text{tok}} \in \mathbb{R}^{V \times d}$ is the token embedding table, $\mathbf{W}_{\text{jamo}} \in \mathbb{R}^{B_j \times d/2}$ and $\mathbf{W}_{\text{char}} \in \mathbb{R}^{B_c \times d/2}$ are hashed embedding tables for jamo 3-grams and characters respectively, and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d/2 \times d}$ projects subword features back to the model dimension. The self-gating mechanism (Eq. 5) amplifies confident subword features while suppressing noise, computed as $\text{SiLU}(\mathbf{f}) \odot \mathbf{f} = \mathbf{f}^2 \odot \sigma(\mathbf{f})$ where σ is the sigmoid function.

⁴Preliminary ablations comparing $n = 3, \{3, 6\}$, and $\{3, 6, 9\}$ found comparable performance with $n = 3$ slightly ahead; a more thorough investigation is left to future work.

5 Results

All results, visualizations, and analyses in this paper are produced from the final supervised-fine-tuned checkpoint. For the KLUE and temporal classification numbers, we additionally apply classifier-head fine-tuning on top of that checkpoint for parity with the encoder-only baseline configurations; all other reported behavior (including the document-level script-invariance metrics, the next-token-prediction probe, and the UMAP visualizations) reflects the supervised-fine-tuned model directly.

5.1 Korean NLU Benchmarks

We evaluate Han2Han on KLUE (Park et al., 2021), a standard Korean NLU benchmark. The encoder-only baseline cluster spans 110M–270M parameters and was pre-trained on 131B to $\sim 6.3T$ tokens (Table 6); the encoder-decoder baseline SKT-KoBART-base-v1 sits at 124M parameters with an undisclosed token budget. Han2Han is comparably sized at 169M parameters but trains on 35B tokens, 4–200 $\times$ fewer than the encoder baselines.

Despite this smaller training budget, Han2Han matches or approaches the baselines on YNAT (Table 1). On STS the reading depends on the split, and we report both. Under a matched fine-tuning loss, Han2Han’s 84.4 sits mid-range among the baselines on the KorSTS dev split; on the held-out test split the Korean-native encoders—all pre-trained on far more contemporary Korean than our 35B mostly historical tokens—lead every Han2Han model, 85.9 for the strongest against 75.1 here. We therefore make no parity claim on contemporary STS. NLI is the weakest dimension, consistent with our pre-training mixture deliberately prioritizing script-invariance and morpheme structure over premise-hypothesis reasoning; Han2Han’s claims live on script-invariance (Section 5.2) and historical text (Section 5.3), where the ordering reverses.

5.2 Script-Invariant Representations

To evaluate script-invariance, we compare encoder-decoder representations across three input configurations: (1) original Hanja article given to both encoder and decoder, (2) original Hanja to encoder with Hangul-transcribed text to decoder (testing cross-attention bridging), and (3) Hangul-transcribed text to both. Document representations are obtained by mean-pooling decoder outputs, which already incorporate encoder informa-

	YNAT	STS		NLI
Model	dev	dev	test	dev
XLM-RoBERTa-base	80.6	81.7	79.0	75.2
KLUE-BERT	84.9	87.7	82.5	80.3
KLUE-RoBERTa-base	84.1	86.8	84.2	83.7
KoELECTRA-base-v3	83.5	88.1	85.9	85.4
SKT-KoBART-base-v1	86.0	84.9	81.4	77.6
Han2Han	84.5	84.4	75.1	74.7

Table 1: KLUE benchmark scores. YNAT: macro F1 $\times 100$, KLUE dev. NLI: accuracy %, KLUE dev. STS: KorSTS Spearman $\times 100$ under a fine-tuning loss matched across all models, reported on both the dev and the held-out test split. KLUE does not release test labels for YNAT and NLI, so those columns are dev throughout.

tion through cross-attention; the same procedure and supervised-fine-tuned checkpoint stage apply to T5Gemma 2 baselines for matched comparison.

The UMAP (McInnes et al., 2018) visualization in Figure 3 provides empirical evidence of script-invariant semantic representations. We evaluate on 1,124 hand-transcribed art criticism articles from the 1920s–1940s CASASIA corpus (Center for Art Studies, Korea, 2016). While T5Gemma 2 (786M parameters) segregates documents entirely by script—creating isolated “islands” for Hanja and Hangul versions of identical content—Han2Han clusters by semantic content with all three script conditions interleaved.

Quantitative document-level metrics (Table 9, Appendix F) confirm what Figure 3 shows visually. T5Gemma 2 without the Han2Han recipe segregates documents entirely by script: bidirectional k -NN of 0.03 means a document’s nearest cross-script neighbor is almost never itself. Applying the recipe to T5Gemma 2 via continued pre-training (CPT) brings k -NN to 0.57 and mean cosine to 0.97; Han2Han 169M trained from scratch reaches k -NN 0.61 with $4.7\times$ fewer parameters and $1/57$ the pre-training tokens. We read this as evidence for the recipe, which improves all three metrics when applied to T5Gemma 2, and not as evidence that the architecture is separately better: Han2Han leads only on the rank-based metric and trails T5Gemma 2 + recipe on cosine (0.84 vs. 0.97) and script mixing (0.78 vs. 0.89). Component attribution is addressed directly by the matched-token ablations of Section 5.5, not by this comparison.

Lexical-level cognate retrieval across seven Korean LMs and the recipe-adapted T5Gemma 2 (Appendix H) shows the same pattern at the word level: only Han2Han 169M and T5Gemma 2 + recipe

(CPT) reach high centered top-1 cross-script retrieval, while existing models—including 1.2–1.5B-parameter decoder-only Korean LMs—remain at or below 0.41. These cross-model comparisons are a capability survey, not an attribution experiment: the models differ in architecture, objective, tokenizer, historical-data exposure, and fine-tuning stage at once. They establish that current Korean-inclusive pre-training does not deliver script-invariance at any scale we could test, and nothing about which of our components is responsible.

5.3 Temporal Classification

Model	Test F1	Coverage
Han2Han	76.0	99.9[†]
T5Gemma 2	54.3	99.9 [†]
KLUE-BERT	54.3	85.4
SKT-KoBART-base-v1	53.6	—
KLUE-RoBERTa-base	50.9	85.4
KoELECTRA-base-v3	49.4	—
XLM-RoBERTa-base	47.5	99.8

Table 2: Temporal classification on the Chojoongdong corpus (1920s–1990s), **article-level held-out test split**, with **all inputs converted to Hangul**—removing script-based cues. Macro F1 $\times 100$. Coverage measures native-vocabulary tokenizer coverage on the original mixed-script text ([†]each tokenizer misses a single token; not measured for SKT-KoBART and KoELECTRA). Parameter counts in Table 6; T5Gemma 2 is 786M. Han2Han and T5Gemma 2 achieve practically identical near-complete coverage yet differ by 21.7 F1 points: tokenizer access is not the operative variable for this task. Each baseline received its own optimizer and learning-rate sweep; every Han2Han model shares a single fine-tuning recipe tuned once on the reference arm’s validation split.

To evaluate Han2Han’s capability for historical text analysis, we test temporal classification on our Chojoongdong newspaper corpus (Chosun Ilbo, JoongAng Ilbo, Dong-a Ilbo). Models classify articles into their year of publication (1920s–1990s), in 5-year buckets, based solely on linguistic features. A key challenge is that Hanja usage declined dramatically across this period (49% in the 1920s to 3% by the 1990s; Table 4), so models could trivially classify by script ratio rather than semantic understanding. To control for this, we convert *all inputs to Hangul* before evaluation—every model receives identical Hangul-only text, forcing reliance on vocabulary, style, and discourse patterns rather than script distribution. Evaluation uses an article-level held-out split: 77,920 articles partitioned 46,752/15,584/15,584 into train, vali-

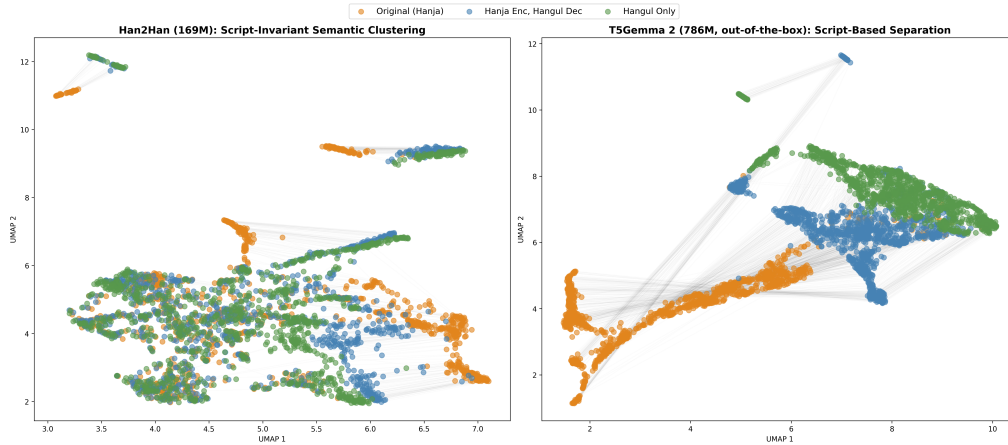


Figure 3: **Empirical evidence of script-invariant representation learning.** UMAP visualization comparing Han2Han 169M (left) against T5Gemma 2 786M (right) on 1,124 articles from the 1920s–1940s CASASIA art-criticism corpus. Each document is processed in three conditions: Hanja→Hanja (orange), Hanja→Hangul via cross-attention (blue), and Hangul→Hangul (green); gray lines connect the three representations of each document. Han2Han interleaves the three conditions within semantic clusters; T5Gemma 2 segregates entirely by script. Quantitative metrics in Section 5.2.

dation, and test across sixteen five-year buckets, with every article registered in a URL exclusion list that the pre-training data loader drops before any training batch is drawn. All figures are test-set macro F1.

Table 2 shows Han2Han outperforms all baselines on this Hangul-only evaluation: 76.0 against 54.3 for the strongest baselines—KLUE-BERT and the $4.7\times$ larger T5Gemma 2 alike—a margin of 21.7 points. Temporal classification here relies on semantic and stylistic features rather than script distribution, and neither parameter count nor tokenizer coverage closes the gap.

Pre-training exposure. The exclusion mechanism above post-dates the released checkpoint, whose hold-out was re-drawn as pre-training entered its final cooldown phase. Some evaluation articles may therefore have received gradients—as unlabeled denoising text, never with temporal labels—during the last 20% of training, and the affected subset cannot be reconstructed. Removing the year-prediction objective, the only channel that consumes year metadata, moves the task by -0.3 points, within noise, and the ablation fleet, which excludes all evaluation articles by construction, reproduces the result at full strength: the 24B-token full-recipe model reaches 76.6 (three-seed mean) against the same 54.3 baseline under the identical protocol. Separately, on the 38 CASASIA documents that a budget truncation dropped before pre-training ever saw them, script-invariance matches a year-matched trained control (k -NN 0.83 in both);

the invariance of Section 5.2 is not a memorization artifact.

Exposure versus recipe. This task measures historical-domain exposure at least as much as it measures the recipe, and is therefore not presented as evidence for the recipe per se; the ablations separate the two. Removing the historical corpora costs 3.55 ± 0.20 points; removing Hanja from the pre-training text entirely (an all-Hangul control trained from scratch) costs 3.69 ± 0.56 . The two removals cost about the same here but are worlds apart on cross-script alignment: without the historical domain the model still retrieves cognates at ten times chance (24.5 centered top-1, chance 2.5), having learned the correspondence from a residue where Hanja is roughly 0.3% of characters, some forty times sparser; without Hanja at all it sits at chance (5.0). Exposure buys the temporal number; what the recipe needs is not the historical domain but the presence of both scripts, in however small a quantity.

5.4 Cross-Script Generation

The results above rest on encoder representations. A decoder-side companion probes Hangul→Hanja reconstruction on a single 1922 *Chosun Ilbo* article, scored against its 531 reference Hanja under alignment-fair matching (Table 10, Appendix G). Off-the-shelf instruction-tuned T5Gemma 2 produces only Hangul echo despite $\sim 2T$ pre-training tokens. After continued pre-training with the Han2Han recipe it recovers 39% of the reference

Hanja; Han2Han, trained from scratch with $4.7\times$ fewer parameters and 1/57 the pre-training tokens, recovers 38% (35% with its 20-character Hanja prefill excluded). Both models fail in the same way, substituting a more frequent reading of a syllable for the contextually correct one (e.g., 朴泳孝 $\rightarrow$ 朴永孝). One article supports no quantitative claim and none is drawn; the cross-script alignment claim is carried by the encoder-side metrics of Sections 5.2 and 5.5.

5.5 Component Attribution

Comparison	Δ F1	Seeds
– Historical corpora (4B)	-3.55 ± 0.20	3/3 neg.
All-Hangul pre-training (4B)	-3.69 ± 0.56	3/3 neg.
Recipe – vanilla UL2 (24B)	$+1.00 \pm 0.33$	3/3 pos.
Recipe – vanilla UL2 (4B)	$+1.14 \pm 0.61$	3/3 pos.
4B $\rightarrow$ 24B tokens	+0.4	overlap

Table 3: Paired three-seed effects on held-out temporal classification (test macro F1 $\times 100$), each against its matched reference arm. “Seeds” reports per-seed sign consistency. The full nine-arm leave-one-out fleet, with absolute scores, is in Appendix D.

To attribute components in the final architecture rather than in an earlier variant, we train matched arms from scratch: nine leave-one-out arms at 4B tokens against a full-recipe reference, plus a full-recipe/vanilla-UL2 pair at 24B. Every arm drops the temporal evaluation articles at the data loader before any split is drawn.

Two effects survive the noise (Table 3). Pre-training data dominates: removing the historical corpora, or Hanja entirely, each costs about 3.6 points, negative in every seed. The recipe is worth about one point at matched tokens, positive in every seed at both budgets, and larger than the +0.4 (three-seed mean) a six-fold token increase buys. The floor is high—the weakest arm’s three-seed mean is still 72.7, eighteen points above the strongest baseline—matching the decomposition of Section 5.3: exposure sets the level, the recipe adds a smaller reliable increment.

The sub-character embeddings do not separate from noise at these budgets: their leave-one-out effects on temporal classification, STS, YNAT, and NLI all sit within run-to-run variation, and on the frozen-encoder cognate probe the arm without them scores *above* the reference. We therefore present jamo and character fusion as an architectural prior rather than a component with a demonstrated score contribution; these budgets may simply be too small to resolve it. KLUE-YNAT is flat

across all ten arms (84.8–85.4) and the recipe’s NLI effect is a three-seed null (-0.02 ± 0.21): the recipe neither helps nor harms contemporary NLU, which is the intended trade.

6 Conclusion

We have demonstrated a structural limitation of distributional semantics: when orthographic variants rarely co-occur in training data, language models do not learn their equivalence by default. Existing models—including T5Gemma 2 (786M parameters), trained on a 2T-token Korean-inclusive UL2 mixture—segregate Hanja and Hangul representations of the same content into disjoint regions of embedding space (Figure 3, Table 9).

The broader implication is that language modeling benefits from language-specificity. Korean’s mixed-script history made this structural problem visible, but the principle generalizes: language structure can be encoded *a priori* rather than attained through learning, and where orthography carries meaning beyond surface form, purpose-built architectures may reach what scale alone cannot.

7 Limitations

Corpus composition: Han2Han’s pre-training corpus deliberately overweights mixed-script historical Korean (Appendix A), which yields the script-invariance and historical-text capabilities we target while reducing coverage on broad-curriculum NLU (Table 1). Additional epochs over the existing corpus—demonstrated by Raffel et al. (2020) to be effective up to 64 repetitions—are a natural direction for further work.

Evaluation benchmarks: Standard Korean NLP benchmarks are designed for contemporary text, not historical data processing. Han2Han’s byte-to-token reconstruction objective contributes to near-complete tokenizer coverage on mixed-script text (Table 2), but KLUE tasks do not directly probe this capability. Development of historical Korean NLP benchmarks remains an important direction, as does targeted evaluation of whether Hanja-grounded pre-training yields measurably better understanding of contemporary Korean text via Sino-Korean semantic structure—e.g., subset analysis of Sino-Korean and loanword categories within CLICK (Kim et al., 2024a), or hand-curated probes testing semantic precision on Sino-Korean compounds.

Korean-specific scope: Han2Han’s recipe en-

codes Korean-specific structural priors (jamo decomposition, MeCab-Ko morpheme boundaries, Hanja↔Hangul transcription) and is therefore not directly transferable to languages without analogous orthographic structure. The Korean copyright constraints discussed in Appendix A.3 also fundamentally limit the volume of contextually-grounded mixed-script training data available to any single model. A natural extension addressing both is multilingual training that shares etymological structure across writing systems—particularly across the broader CJK family, whose shared logographic substrate makes the same shared-sub-character principle directly applicable. We leave cross-linguistic evaluation to future work.

Ethics Statement

We adhere to the ethical guidelines outlined in the ACL Code of Ethics. Beyond this baseline, we identify three considerations relevant to downstream use of Han2Han in humanities and historiographic research.

Mis-transcription and scholarly validation. Han2Han’s dictionary-based Hanja→Hangul transcription module is high-coverage but not infallible: rare classical compounds, proper nouns with non-standard readings, and historically variant orthographies may be transcribed incorrectly. Outputs intended for scholarly use should be cross-checked against primary sources rather than treated as ground truth, particularly for historiographic claims dependent on precise transcription.

Research aid, not philological substitute. Han2Han is intended to support, not replace, expert philological analysis of historical Korean texts. Researchers using model-assisted transcription or classification in published scholarship should disclose model use and validate interpretive claims independently.

Reproducibility under Korean copyright law. Korean copyright restrictions on the redistribution of large mixed-script training corpora (Appendix A.3) create a structural tension between reproducible research and IP compliance that the Korean NLP community has not yet collectively resolved. We document our data sources to support independent reconstruction, and release our pre-training, fine-tuning, and evaluation code, the model weights, and the temporal evaluation splits under the Apache 2.0 license at <https://github.com/cadazar/han2han>; the un-

derlying newspaper corpus itself cannot be redistributed. We encourage the field to develop shared mechanisms—preprocessed weights, agreed-on evaluation slices, or new licensing frameworks—that enable reproducibility without violating copyright restrictions.

Acknowledgements

This research was supported by BK21 FOUR (Fostering Outstanding Universities for Research, No. 2120240915469); Culture, Sports and Tourism R&D Program, funded by the Ministry of Culture, Sports and Tourism, through the Korea Culture Technology Planning and Evaluation Institute, an affiliated institute of the Korea Creative Content Agency in 2026 (Project Number: RS-2026-25506607). We also thank the TPU Research Cloud for generously providing access to TPU VMs for our research.

References

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Center for Art Studies, Korea. 2016. [Joseon art exhibition database \(조선미술전람회\)](#). CASASIA Digital Archive. Accessed: 2025.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*, 25(70):1–53.
- Marco Cignetta, Sangwhan Moon, Lawrence Wolfsonkin, and Naoaki Okazaki. 2023. [Parameter-efficient Korean character-level language modeling](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2350–2356, Dubrovnik, Croatia. Association for Computational Linguistics.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. [Word translation without parallel data](#). In *International Conference on Learning Representations (ICLR)*.

- Aleksandr Dremov, Alexander Hägele, Atli Kosson, and Martin Jaggi. 2025. [Training dynamics of the cooldown stage in warmup-stable-decay learning rate scheduler](#). *Preprint*, arXiv:2508.01483.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. 2025. Metadata conditioning accelerates language model pre-training. *arXiv preprint arXiv:2501.01956*.
- Gemma Team. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. [KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.
- Jianfei He, Shichao Sun, Xiaohua Jia, and Wenjie Li. 2024. [Recovery should never deviate from ground truth: Mitigating exposure bias in neural machine translation](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 68–79, Sheffield, UK. European Association for Machine Translation (EAMT).
- Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. 2020. [Query-key normalization for transformers](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, Online. Association for Computational Linguistics.
- Taehee Jeon, Bongseok Yang, Changhwan Kim, and Yoonseob Lim. 2023. [Improving Korean NLP tasks with linguistically informed subword tokenization and sub-character decomposition](#). *Preprint*, arXiv:2311.03928.
- Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. 2024. [Muon: An optimizer for hidden layers in neural networks](#).
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. [CLiCK: A benchmark dataset of cultural and linguistic intelligence in Korean](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.
- 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.
- SungHo Kim, Juhyeong Park, Yeachan Kim, and SangKeun Lee. 2024b. [KOMBO: Korean character representations based on the combination rules of subcharacters](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5102–5119, Bangkok, Thailand. Association for Computational Linguistics.
- Junyoung Lee, Marco Cognetta, Sangwhan Moon, and Naoaki Okazaki. 2025. [Jamo-level subword tokenization in low-resource Korean machine translation](#). In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 66–80, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Qilong Li and Chongsheng Zhang. 2025. [UCR: A unified character-radical dual-supervision framework for accurate Chinese character recognition](#). *Pattern Recognition*, 162:111373.
- Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. [Confidence-aware scheduled sampling for neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2327–2337, Online. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. [The flan collection: Designing data and methods for effective instruction tuning](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.
- Leland McInnes, John Healy, and James Melville. 2018. [UMAP: Uniform manifold approximation and projection for dimension reduction](#). *arXiv preprint arXiv:1802.03426*.

- Eunjeong L. Park and Sungzoon Cho. 2014. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea.
- Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. [Subword-level word vector representations for Korean](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2429–2438, Melbourne, Australia. Association for Computational Linguistics.
- Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. [KLUE: Korean language understanding evaluation](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. 2019. [Additive compositionality of word vectors](#). In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 387–396, Hong Kong, China. Association for Computational Linguistics.
- Noam Shazeer. 2020. [GLU variants improve Transformer](#). *Preprint*, arXiv:2002.05202.
- 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.
- Seyoung Song, Nawon Kim, Songeun Chae, Kiwoong Park, Jiho Jin, Haneul Yoo, Kyunghyun Cho, and Alice Oh. 2025. [Open Korean Historical Corpus: A millennia-scale diachronic collection of public domain texts](#). *Preprint*, arXiv:2510.24541.
- Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2022. [UL2: Unifying language learning paradigms](#). *Preprint*, arXiv:2205.05131.
- Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. [Foundation transformers](#). *Preprint*, arXiv:2210.06423.
- Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. [TSDAE: Using Transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning](#). *Preprint*, arXiv:2104.06979.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [ByT5: Towards a token-free future with pre-trained byte-to-byte models](#). *Transactions of the Association for Computational Linguistics*, 10:291–306.
- 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: Pre-trained 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.
- Kang Min Yoo, Taeuk Kim, and Sang-goo Lee. 2019. [Don’t just scratch the surface: Enhancing word representations for Korean with hanja](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3528–3533, Hong Kong, China. Association for Computational Linguistics.
- Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025a. [Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation](#). *Preprint*, arXiv:2504.06225.
- Biao Zhang, Paul Suganthan, Gaël Liu, Ilya Philipov, Sahil Dua, Ben Hora, Kat Black, Gus Martins, Omar Sanseviero, Shreya Pathak, Cassidy Hardin, Francesco Visin, Jiageng Zhang, Kathleen Kenealy, Qin Yin, Xiaodan Song, Olivier Lacombe, Armand Joulin, Tris Warkentin, and Adam Roberts. 2025b. [T5Gemma 2: Seeing, reading, and understanding longer](#). *Preprint*, arXiv:2512.14856.
- Jack Zhang, Noah Amsel, Berlin Chen, and Tri Dao. 2026. [Gram Newton-Schulz: A fast, hardware-aware Newton-Schulz algorithm for Muon](#).

A Training Corpus Details

A.1 Twentieth Century Newspaper Corpus

The core historical component of the training corpus was collected from over seven million newspaper articles written and published between 1920 and 1999 in the South Korean newspapers *Chosun Ilbo*, *Dong-a Ilbo*, and *JoongAng* (collectively referred to as Chojoongdong).⁵ All texts were digitized versions of the original printed material, so there were inevitably several transcription mistakes.⁶ The title, author, newspaper section, publication date, and URL were all collected along with the texts. All materials are available directly on the websites of the respective news agencies but cannot be re-distributed according to South Korean law. In any case, they are fairly easily accessible and their collection can be automated.

Periods	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	Total
Number of Documents	718K	1.0M	185K	470K	758K	959K	1.1M	1.9M	7.1M
Number of Tokens	144M	245M	39M	107M	177M	248M	307M	516M	1.74B
Avg. Tokens Per Article	200	234	211	227	233	259	286	276	241
Hanja ratio	0.49	0.44	0.42	0.38	0.27	0.13	0.07	0.03	0.28

Table 4: Statistics of the Chojoongdong newspaper corpus (*Chosun Ilbo*, *Dong-a Ilbo*, and *JoongAng*). Tokens are counted according to the SentencePiece tokenization scheme. The declining Hanja ratio reflects the gradual transition from mixed-script to Hangul-only publishing across the 20th century.

A.2 Additional Data Sources

Beyond the historical newspaper corpus (1.74B tokens), we assembled additional Korean training data. This includes: (1) **Specialized Domain Corpus** (전문분야 말뭉치) comprising patents, legal documents, court cases, and academic papers; (2) **Web/Encyclopedia** sources including OSCAR, NamuWiki, Korean Wikipedia, and the Korean subset of FinePDFs; (3) **Modu Corpus** (모두의 말뭉치) spanning written, spoken, news, and messenger text from the National Institute of Korean Language; (4) **Book Corpus** from the National Library of Korea across 74 KDC categories; (5) **Korean Petitions** from the Blue House; and (6) the **Open Korean Historical Corpus** (OKHC) (Song et al., 2025), covering a millennium of premodern and early-modern Korean text. Data sources are weighted during training, with curated historical data prioritized over web crawls.

⁵All romanization of Korean terms in this paper follows the McCune-Reischauer standard, common terms with established romanization conventions omitted.

⁶For example, in the *Dong-a Ilbo* articles, the character 象 (*sang*, “form”), as in 抽象 (*ch’usang*, “abstract”) was repeatedly missing, likely due to encoding inconsistencies. However, many of these errors were either consistent enough to be corrected in preprocessing or inconsistent enough to have little effect on the overall corpus.

Category	Documents	Tokens
<i>Specialized Domain (전문분야 말뭉치)</i>		
Patents (특허)	543K	3.94B
Laws (법령)	99K	211M
Local Regulations (자치법규)	110K	181M
Court Cases (판례)	76K	162M
Academic Papers (논문)	233K	70M
Bills (의안)	62K	22M
<i>Web/Encyclopedia</i>		
FinePDFs (Korean)	1.09M	12.30B
OSCAR (Korean)	3.68M	2.35B
NamuWiki	568K	1.22B
Wikipedia (Korean)	643K	289M
<i>Modu Corpus (모두의 말뭉치)</i>		
Written	10K	578M
News	1.02M	456M
Spoken	7.41M	270M
Messenger	4K	6M
<i>Historical and Other</i>		
OKHC [†]	7.42M	1.63B
Book Corpus (74 KDC)	584K	493M
Blue House Petitions	422K	99M
Classical Texts	13K	2M
Subtotal (itemized above)	24.0M	24.28B

Table 5: Additional training corpus statistics. Combined with the Chojoongdong newspaper corpus (Table 4, 1.74B tokens), the pre-training mixture comprises approximately 26B unique tokens. Data sources are weighted during training, with curated historical data prioritized over web crawls. [†]OKHC contains 17.88M documents in total; 58.5% have their text content removed due to copyright restrictions (see Appendix A.3), leaving 7.42M documents with available text contributing the reported token count.

A.3 Data Licensing and Availability

Training Korean language models faces systemic challenges related to intellectual property law. While large-scale Korean text corpora exist, redistribution restrictions under Korean copyright law limit their accessibility for reproducible research.

The Open Korean Historical Corpus (Song et al., 2025) represents a valuable effort to provide accessible historical Korean text data. However, we note that 58.5% of documents in OKHC have their text content removed due to copyright restrictions, with only metadata available—a limitation the corpus authors explicitly acknowledge as resulting from Korean IP law rather than curatorial choice.

This situation reflects a broader challenge in Korean NLP: the tension between reproducible research (requiring redistributable data) and copyright compliance. Our historical newspaper corpus faces similar constraints—while the collection process is automatable from public archives, redistribution is prohibited under South Korean law. We encourage future work on developing mechanisms for sharing preprocessed model weights or evaluation splits that enable reproducibility without violating copyright restrictions.

B Hanja-Aware Morpheme Tokenization

美術展覽會 → 미술전람회 →
 ㅁ ㅣ -ㅅ ㅌ ㄹ ㅅ ㅌ ㅁ ㅎ ㅌ - →
 ㅁ ㅣ -, ㅣ -ㅅ, -ㅅ ㅌ, ㅅ ㅌ ㄹ, ㅌ ㄹ ㅅ,
 ㄹ ㅅ ㅌ, ㅅ ㅌ ㄹ, ㅌ ㄹ ㅅ, ㄹ ㅌ ㅁ,
 ㅌ ㅁ ㅎ, ㅁ ㅎ ㅌ, ㅎ ㅌ -

Figure 4: Jamo 3-gram extraction for “art exhibition.” Hanja is first transcribed to Hangul, then decomposed into jamo (onset-nucleus-coda), with – marking empty codas. A sliding window extracts overlapping 3-grams, each spanning 3 input symbols but 2–3 phonetic jamo positions because – placeholders are not phonetic units. 3-grams are hashed to embedding buckets following Bojanowski et al. (2017). Because transcription precedes decomposition, 美術展覽會 and 미술전람회 produce identical jamo features. This is the mechanism that enables script-invariant representations.

A key challenge in applying morpheme-aware corruption to mixed-script Korean text is that standard Korean morphological analyzers like MeCab-ko (Park and Cho, 2014) are trained primarily on Hangul text. When processing Hanja characters directly, MeCab-ko often produces inconsistent or incorrect morpheme boundaries:

```
>>> mecab.pos("美術展覽會")      # Hanja input
      [ ('美術展', 'NNP'),          # INCORRECT: splits mid-word
        ('覽', 'SH'),
        ('會', 'NNG')]

>>> mecab.pos("미술전람회")      # Hangul input
      [ ('미술', 'NNG'),            # CORRECT: proper boundaries
        ('전람회', 'NNG')]
```

As shown above, the Hanja word 美術展覽會 (*misuljeollamhoe*, “art exhibition”) is incorrectly split as 美術展/覽/會, while its Hangul equivalent 미술전람회 correctly segments as 미술 (“art”) + 전람회 (“exhibition”). This inconsistency arises from the limited coverage of Hanja in MeCab-ko’s dictionary.

To ensure script-invariant morpheme boundaries, we implement a **Hanja-aware morpheme tokenization** pipeline:

1. **Transcribe**: Convert input text to Hangul using our transcription module (preserving 1:1 character mapping)
2. **Analyze**: Apply MeCab-ko morphological analysis to the Hangul text
3. **Map back**: Use character positions to extract corresponding segments from the *original* text, preserving Hanja characters

```
>>> hanja_aware_tokenize("美術展覽會")
      ['美術', '展覽會']          # Correct boundaries, original script
```

This approach guarantees that morpheme-aware corruption operates on linguistically meaningful units regardless of script, which is essential for our bidirectional transcription training where the same semantic content may appear in either Hanja or Hangul form.

C Architecture and Training Details

Model	Params	Train Tokens
XLM-RoBERTa-base	270M	~6.3T (est.)
KLUE-BERT	110M	~131B
KLUE-RoBERTa-base	110M	~1T
KoELECTRA-base-v3	110M	~197B
SKT-KoBART-base-v1	124M	undisclosed
Han2Han	169M	35B

Table 6: Baseline parameter and training-token counts. Train Tokens computed as training steps $\times$ batch size $\times$ sequence length where available; XLM-RoBERTa estimated from large model specifications.

Layers and parameter sharing. Han2Han has 18 encoder layers, 18 decoder layers, hidden dimension 640, and FFN dimension 2048; the depth and dimension ratio follow Gemma 3’s smallest configuration (Gemma Team, 2025), while the overall encoder-decoder structure is closest to T5Gemma 2 (Zhang et al., 2025b). We tie the main weights of the encoder and decoder while keeping biases, RMSNorm scales, subtoken linear projections, and subtoken embeddings untied for per-component expressivity; sub-layer normalization (SubLN) (Wang et al., 2022) recovers the representational capacity that fully-shared main weights would otherwise sacrifice.

Attention pattern. Self-attention uses multi-query attention (MQA: 4 query heads, 1 KV head) for memory efficiency; cross-attention uses GQA-2 (4 query heads, 2 KV heads) on a dedicated projection set. This contrasts with T5Gemma 2 (Zhang et al., 2025b), which uses a "merged" attention scheme where decoder self-attention and cross-attention share projections by concatenating encoder hidden states onto decoder hidden states with autoregressive masking. Han2Han keeps cross-attention separate because it

is the primary semantic bridge between scripts—Hanja-encoded input conditioned on Hangul-decoded output and vice versa—and the recipe’s emphasis on script-bridging justifies the small parameter cost. Following Gemma 3 (Gemma Team, 2025), attention alternates 5:1 between sliding-window layers (window 128, RoPE $\theta = 10^4$) and full-attention layers (RoPE $\theta = 5 \times 10^5$) across encoder, decoder, and decoder cross-attention. Cross-attention does not use positional embeddings. QK normalization (Henry et al., 2020) is applied throughout, and feed-forward blocks use GeGLU (Shazeer, 2020).

Script tokens. We add two special document-level tokens, `<hangul>` and `<hanja>`, prepended to every sequence—analogous in mechanism to BERT’s `[CLS]` or mBART’s `<lang>` markers (Liu et al., 2020), but signaling the *script* of the input rather than language or sequence position. Unlike mBART’s language tokens, Han2Han’s script tokens mark orthographically different but semantically equivalent representations of the same Korean content, enabling mismatched markers in the encoder and decoder during pre-training where mBART performed exclusive single-marker pre-training. Hanja→Hangul transcription is well-specified at the compound level: our transcription module resolves readings through a curated compound dictionary ($\sim 32\text{K}$ compound entries with longest-match-first lookup) backed by $\sim 27\text{K}$ single-character defaults, producing the context-appropriate readings illustrated in Section 2.1. The reverse direction, Hangul→Hanja, is one-to-many (cf. the $\uparrow$ example in Section 2.1), and this one-to-many disambiguation is the learning problem that the bidirectional transcription objective targets. Including the script tokens and decomposed jamo letters during pre-training signals the model to capture paired Hanja–Hangul sequences and the appropriate jamo composition for each Sino-Korean character’s context-dependent readings.

Optimizer and training. Han2Han trains in BF16 throughout (parameters and optimizer state) on the 35B-token corpus described in Section 3. We use Muon (Jordan et al., 2024) with Gram Newton-Schulz orthogonalization (Zhang et al., 2026). Learning rate is 2.66×10^{-4} on a 10% warmup / 70% constant / 20% square-root cooldown schedule, following the cooldown-shape analysis of Dremov et al. (2025); effective batch size is approximately 4.8M tokens per update (counting packed encoder and decoder tokens, including sentinels), with weight decay 0.1. We perform gradient norm clipping at global norm 1.0, applied only to the AdamW arm of the optimizer. Weights are initialized with truncated-normal fan-in variance-scaling at scale 0.1. Consistent with the single-epoch / no-repetition regime, we apply no LayerDrop, residual or attention dropout, or label smoothing. Pre-training ran on a TPU v5e-64 pod slice for approximately 35 hours; supervised fine-tuning ran on a single TPU v6e-8 host for approximately 9 hours, using the same 2048-token sequence length and sequence packing (across both encoder and decoder) as pre-training. Full hyperparameters are listed in Table 7.

Training Configuration	
Optimizer	Muon + Gram Newton-Schulz
Muon momentum β	0.95
AdamW arm (1D params) β_1 / β_2	0.90 / 0.95
Weight decay	0.1
Learning rate	2.66×10^{-4}
LR schedule	10% warmup / 70% constant / 20% sqrt cooldown
Effective batch size	$\sim 4.8\text{M}$ tokens / update (packed)
Sequence length	2048 tokens
Gradient clipping	1.0 (global norm, AdamW arm only)
Training duration	35B tokens
Precision	BF16 (model, params, optimizer state)
Initialization	
Kernel init	Variance-scaling (fan-in, truncated normal), scale 0.1
Compute	
Pre-training hardware	TPU v5e-64 pod slice
Pre-training wall-clock	~ 35 hours
SFT hardware	TPU v6e-8 single host
SFT wall-clock	~ 9 hours
Regularization (all disabled)	
LayerDrop	0
Residual / attention dropout	0
Label smoothing	0

Table 7: Training hyperparameters for Han2Han. The single-epoch / no-repetition regime (Section 3) motivates disabling all forms of stochastic regularization; representational priors are provided by the data and objective rather than by dropout.

D Ablation Study Results

Table 8 reports matched ablation arms trained from scratch in the final architecture: nine leave-one-out arms at 4B tokens against a full-recipe reference, plus a full-recipe/vanilla-UL2 pair at 24B tokens. Every arm excludes the temporal evaluation articles from its pre-training data by URL, so no arm trains on evaluation text.

Arm	Tokens	Temporal (test F1)		Cognate top-1		YNAT
		abs.	Δ	abs.	Δ	abs.
Full recipe (reference)	4B	75.8	—	72.8	—	85.3
– Jamo embeddings	4B	76.2	+0.4	72.5	−0.3	85.3
– Character embeddings	4B	75.5	−0.3	69.0	−3.8	85.2
– Both sub-token embeddings	4B	75.6	−0.3	82.3	+9.5	84.8
– Morpheme-aware masking	4B	76.2	+0.3	68.0	−4.8	85.3
– Transcription objective	4B	75.7	−0.1	63.5	−9.3	84.9
– Temporal objective	4B	75.5	−0.3	73.8	+1.0	84.9
– Historical corpora	4B	72.5	−3.3	24.5	−48.3	85.4
All-Hangul (single-script input)	4B	72.8	−3.1	5.0	−67.8	84.9
Vanilla UL2 (recipe off)	4B	75.4	−0.5	77.5	+4.8	85.2
Full recipe	24B	76.9	—	83.0	—	86.4
Vanilla UL2 (recipe off)	24B	75.6	−1.3	81.0	−2.0	85.8

Table 8: Leave-one-out ablation arms in the final architecture, seed 42, absolute scores with deltas from the matched reference arm. Temporal classification is test macro F1 $\times 100$ on the held-out split of Section 5.3; cognate retrieval is centered top-1 Hanja→Hangul on a frozen encoder ($\times 100$, chance 2.5, 40 pairs), a representation-level probe rather than a fine-tuned score; YNAT is KLUE dev macro F1 $\times 100$. Seven arms were rerun at seeds 42/43/44; their seed-to-seed standard deviation on temporal classification ranges from 0.12 to 0.36 points, so single-seed deltas below roughly one point should not be read as effects. Paired three-seed effects are reported in Table 3.

Temporal classification. Three effects exceed the seed-to-seed spread: removing the historical corpora (-3.3 at seed 42; -3.55 ± 0.20 over three seeds, negative in every seed), the all-Hangul control (-3.1 ; -3.69 ± 0.56 , negative in every seed), and the full recipe against vanilla UL2 at 24B ($+1.3$ at seed 42; $+1.00 \pm 0.33$ over three seeds, positive in every seed). All other single-arm deltas are smaller than the seed spread.

Sub-token embeddings. The jamo and character arms do not separate from noise: their temporal deltas are within the seed spread, their YNAT deltas are within the ± 0.9 -point binomial band of the KLUE dev set, and on the cognate probe the arm with both sub-token embeddings removed scores 9.5 points above the reference. At these budgets the sub-character embeddings have no demonstrated independent score contribution and are presented as an architectural prior. Their motivation—transcription before decomposition projects Hanja and Hangul cognates onto identical jamo features (Appendix B)—is structural; testing it would need budgets larger than 4B tokens.

Corpus composition. The two data controls measure different things. In the historical newspaper corpus, 13.9% of characters are Hanja and 81% of articles contain at least one (68% contain ten or more). In the remaining data the no-historical-data arm trains on, Hanja falls to roughly 0.33% of characters—about forty times sparser—appearing mainly as isolated glosses in literature, legislation, and academic prose rather than as running text. That arm is therefore a no-domain control rather than a Hanja-free one; the all-Hangul arm is the Hanja-free control. The cognate column distinguishes the two (24.5 versus 5.0 against a chance rate of 2.5) where the temporal column does not.

Contemporary NLU. The recipe is neutral on KLUE. YNAT is flat across the 4B arms (84.8–85.4, against a baseline range of 80.6–86.0 in Table 1), and the recipe/vanilla-UL2 NLI difference is null over three seeds (-0.02 ± 0.21 accuracy, sign-flipping across seeds). Per-arm STS deltas are not individually interpretable: six of nine flip sign between the two fine-tuning protocols, a spread of roughly ± 1.5 points that exceeds every per-arm effect; the overall level is flat across arms.

E Hyperparameter Sweep Analysis

The primary configuration choice informing the final model was the directionality of the Han2Han transcription objective. A short sweep over morpheme-masking ratios, span-length parameters, and transcription directionality on 150M-token training runs found that **bidirectional** Hanja $\leftrightarrow$ Hangul transcription consistently outperformed unidirectional Hanja $\rightarrow$ Hangul on our downstream evaluations.

The remaining denoising-mixture parameters in the final model (R/X/S denoiser ratios and within-denoiser corruption ratios) follow UL2 (Tay et al., 2022) defaults; the morpheme-aware variant within the R-denoiser is described in Section 3. The final model uses morpheme-masking ratio 0.7 (fraction of R-denoiser sequences using morpheme-aware spans rather than token-based Poisson spans) and bidirectional transcription ratio 0.6 (fraction of Hanja-containing sequences for which the transcription objective is applied). Performance was not strongly sensitive to small deviations from these values, consistent with the dominant signal coming from the choice of denoising objective rather than precise mixing ratios.

F Document-Level Script-Invariance Analysis

Table 9 reports the document-level script-invariance metrics for Han2Han 169M, T5Gemma 2 786M, and T5Gemma 2 786M after continued pre-training (CPT) with the Han2Han recipe, evaluated on 1,124 hand-transcribed art-criticism articles from the 1920s–1940s CASASIA corpus (Center for Art Studies, Korea, 2016). Figure 5 pairs the post-CPT UMAP visualization for T5Gemma 2 against Han2Han 169M trained from scratch: visible script-based separation is reduced relative to the pre-adaptation figure (Figure 3), while residual cluster structure remains visible. The high-dimensional metrics confirm that script-invariance is improved (bidirectional k-NN of 0.57, mean cosine of 0.97), so we treat the metrics as the primary evidence and the visualization as qualitative support. Notably, Han2Han 169M slightly exceeds T5Gemma 2 + recipe (CPT) on the rank-based k-NN metric (0.61 vs. 0.57) despite being $4.7\times$ smaller and trained from scratch on $1/57$ the pre-training tokens, indicating that the architecture contributes alongside the recipe to the retrieval geometry.

Model	Params	Cos $\uparrow$	Script Mixing $\uparrow$	k-NN $\uparrow$
T5Gemma 2	786M	0.84	0.49	0.03
T5Gemma 2 + recipe (CPT)	786M	0.97	0.89	0.57
Han2Han (ours)	169M	0.84	0.78	0.61

Table 9: Document-level script-invariance on 1,124 hand-transcribed art-criticism articles (CASASIA, 1920s–1940s). **Cos**: mean cosine similarity between same-document Hanja and Hangul embeddings (higher = aligned across scripts). **Script Mixing**: 1 – silhouette score with script as cluster label (higher = scripts interleaved in embedding space). **k-NN**: bidirectional 1-NN cross-script retrieval accuracy (higher = each document’s nearest cross-script neighbor is itself). Han2Han 169M and T5Gemma 2 786M continually pre-trained with the Han2Han recipe (CPT) achieve comparable script-invariance; T5Gemma 2 without the recipe segregates documents entirely by script.

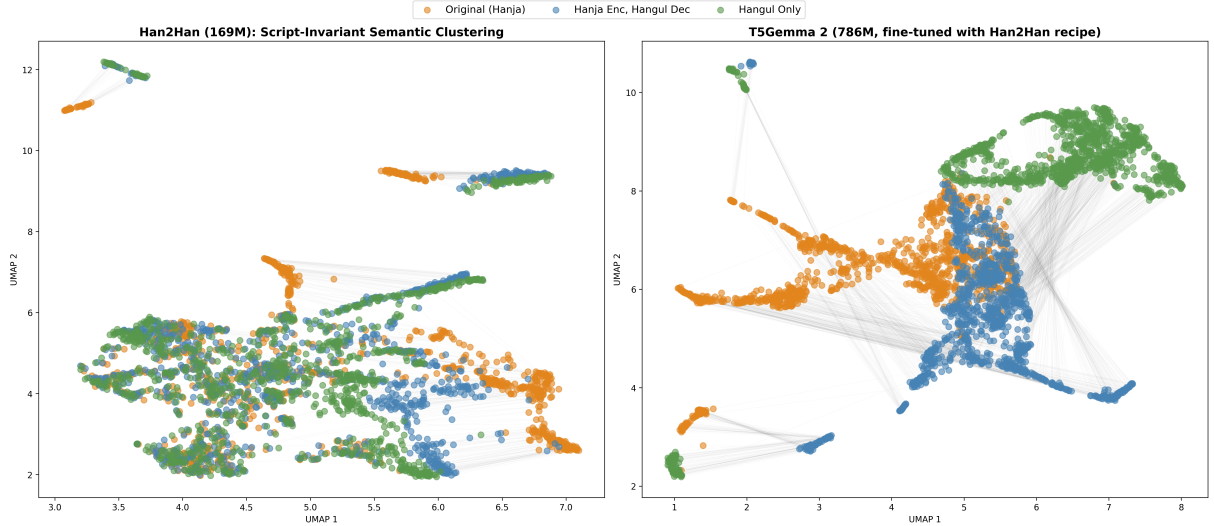


Figure 5: UMAP visualization after applying the Han2Han training recipe to T5Gemma 2 (786M) via continued pre-training, compared against Han2Han 169M trained from scratch. Visible script-based separation is reduced relative to the pre-adaptation figure (Figure 3), though residual cluster structure remains visible. The document-level metrics in Table 9 confirm that script-invariance is improved.

G Cross-Script Generation: Methodology and Qualitative Analysis

Model	Params	Recall	excl. prefill
T5Gemma 2 (IT, no recipe)	786M	0.00*	—
T5Gemma 2 + recipe (CPT)	786M	0.39	—
Han2Han (ours)	169M	0.38	0.35

Table 10: Hangul→Hanja reconstruction recall on a 1922 *Chosun Ilbo* article (531 reference Hanja, alignment-fair scoring). Han2Han decodes from a 20-character Hanja prefill; the last column scores only its generated continuation against the 511 reference Hanja that follow the prefill. The T5Gemma 2 rows use no prefill. *T5Gemma 2 instruction-tuned without the Han2Han recipe produces only Hangul echo (zero Hanja generated).

Setup. A 1922 *Chosun Ilbo* article containing 531 Hanja characters in its original mixed-script form is transcribed to Hangul-only and given to the encoder. The model is asked to reconstruct the original mixed-script form. Scoring is alignment-fair: we align the output with the reference at character level by longest-common-subsequence matching and count the reference Hanja that fall inside matched blocks, so that an insertion or deletion shifts the alignment without penalizing every character downstream of it (e.g., Han2Han produces 第二回 議室 where the reference has 第二會議室).

Decoding. Han2Han uses beam search (beam=8) seeded with the first 20 Hanja of the reference as a decoder prefill. Its generated continuation scores 0.35 against the 511 reference Hanja that follow the prefill; adding the 20 prefilled characters back gives the 0.38 reported in Table 10. T5Gemma 2 + recipe (CPT) decodes greedily with no prefill. Without the prefill, Han2Han’s greedy and beam decodes echo

the Hangul input while the Hanja reading of each syllable sits in the decoder’s top- k ; sampling at $T=0.8$ surfaces it, producing majority-Hanja completions in roughly a quarter of draws. The prefill therefore selects the Hanja branch the decoder already carries rather than supplying capacity it lacks; the cross-script alignment claim itself is tested by the encoder-side metrics in Section 5.2 and Appendix H, which involve no decoding.

Error analysis. Both models fail in the same way. Where they miss, they substitute a more frequent Hanja reading of the syllable for the contextually correct one, most visibly on personal names and rare compounds. Substitutions made by both models: 倣→訪, 朴泳孝→朴永孝, 侯→後, 子→者, 金圭鎮→金奎鎮. By T5Gemma 2 + recipe only: 事案→私案, 初次→抄次, 裨補→備保, 丁大有→丁大裕, 高木背水→高木培秀 (the first two fall inside Han2Han’s prefill). By Han2Han only: 裨補→備報, 帝國美術院→帝國美術園, 第二會議室→第二回議室, 閔丙奭→閔丙錫, 丁大有→正大裕, 李道榮→李道英, 高木背水→高木培. The two outputs differ in alignment rather than in error type: T5Gemma 2 + recipe stays in step with the reference (exact-position Hanja accuracy 0.64), whereas Han2Han’s output drifts through insertions and deletions (0.20), which is why the comparison uses alignment-fair scoring.

Scope. This is a single-article probe, and none of the paper’s claims rest on it. It shows that a model carrying the Han2Han recipe produces contextually plausible Hanja from Hangul input where instruction-tuned T5Gemma 2 without the recipe produces none, and that Han2Han does so at $4.7\times$ fewer parameters. A multi-article evaluation of cross-script generation is future work.

H Cross-Model Cognate Retrieval

To test whether cross-script alignment emerges with scale across the broader Korean LM landscape, we evaluate centered top-1 Hanja→Hangul retrieval on a 40-pair cognate set spanning four lexical frequency quartiles. Given a Hanja word’s embedding, the model must identify its Hangul equivalent among the 40 candidates. We report centered top-1 (cosine similarity after corpus-mean centering) because raw cosine on transformer encoders is dominated by anisotropy (Ethayarajh, 2019), and because rank-based retrieval is rotation- and translation-invariant—the cleanest comparison across architectures with very different absolute geometries (Conneau et al., 2018). Pooling is chosen per architecture: encoder mean-pool for encoder-decoder and encoder-only models; last-token for decoder-only LMs (the only position with full input context under causal masking). Each row of Table 11 is the mean over 10 seeded shuffles.

Model	Params	Centered Top-1
HyperCLOVAX-SEED	1.5B	0.09
KoELECTRA	110M	0.10
XLM-RoBERTa	278M	0.13
T5Gemma 2 (base)	786M	0.21
EXAONE 4.0	1.2B	0.34
KLUE-BERT	110M	0.41
Han2Han (ours)	169M	0.87
T5Gemma 2 + Han2Han recipe (CPT)	786M	0.86

Table 11: Centered top-1 Hanja→Hangul cognate retrieval across Korean LMs, mean over 10 seeded shuffles of a 40-pair cognate set. Models above the line are from-scratch; T5Gemma 2 + Han2Han recipe (CPT; continued pre-training) is reported separately below for direct comparison with Han2Han. The two largest Korean LMs tested (HyperCLOVAX-SEED 1.5B, EXAONE 4.0 1.2B; 7–10 $\times$ Han2Han’s parameter count) sit at the bottom and middle of the column. KLUE-BERT and EXAONE land above their scale peers, plausibly reflecting Hanja-inclusive subsets of their pre-training data; we report them transparently rather than excluding them as outliers.

The two largest Korean LMs in the table (HyperCLOVAX-SEED 1.5B and EXAONE 4.0 1.2B) achieve centered top-1 of 0.09 and 0.34 respectively, despite 7–10 $\times$ Han2Han’s parameter count. Han2Han 169M’s 0.87 is matched only by T5Gemma 2 continually pre-trained with the Han2Han recipe (0.86), confirming at the lexical level what Table 9 shows at the document level: the recipe is the operative variable for cross-script alignment, and Han2Han’s architecture delivers it without requiring a $4.7\times$ -larger base.