Beyond Distribution: Investigating Language Models' Understanding of Sino-Korean Morphemes

Taehee Jeon
Institute for Digital HUSS, Korea University, Seoul, South Korea
taeheejeon22@korea.ac.kr

Abstract

We investigate whether Transformer-based language models, trained solely on Hangul text, can learn the compositional morphology of Sino-Korean (SK) morphemes, which are fundamental to Korean vocabulary. Using $\text{BERT}_{\text{BASE}}$ and fastText, we conduct controlled experiments with target words and their "real" vs. "fake" neighbors—pairs that share a Hangul syllable representing the same SK morpheme vs. those that share only the Hangul syllable. Our results show that while both models—especially BERT—distinguish real and fake pairs to some extent, their performance is primarily driven by the frequency of each experimental word rather than a true understanding of SK morphemes. These findings highlight the limits of distributional learning for morpheme-level understanding and emphasize the need for explicit morphological modeling or Hanja-aware strategies to improve semantic representation in Korean language models. Our dataset and analysis code are available at: https://github.com/taeheejeon22/ko-skmorph-lm.

1 Introduction

Language models (LMs) using distributional information have achieved considerable success, from traditional word embeddings to Transformer-based architectures (Vaswani, 2017). While distribution is essential for semantic learning, human language processing also relies on additional resources beyond distribution—such as phonological and morphological information. These resources differ across languages: for example, understanding Sino-Korean (SK) morphemes[1] is critical for lexical semantics in Korean.

While SK words account for over 57% of the Korean lexicon (Choo and O'Grady, 1996), native speakers often infer word meanings through SK morphemes without explicit instruction in Hanja (Korean Chinese characters). For example (Figure 1), mathematics learners unfamiliar with 진분수 (眞分數 *jin-bun-su* 'proper fraction', where the numerator is smaller than the denominator) or 가분수 (假分數 *ga-bun-su* 'improper fraction', where the numerator is larger) can infer their meanings using SK morpheme knowledge rather than context. Even without prior exposure, knowing morphemes like 진 (眞 *jin* 'true') and 가 (假 *ga* 'false') enables semantic inference.

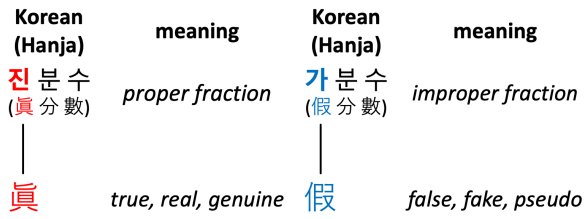

Figure 1: Illustration of how SK morphemes facilitate the interpretation of words like 진분수 (眞分數 *jin-bun-su* 'proper fraction') and 가분수 (假分數 *ga-bun-su* 'improper fraction').

Psycholinguistic studies show that SK morpheme comprehension plays a crucial role in Korean word recognition (Yi and Yi, 1999; Yi et al., 2007; Yi, 2009; Bae et al., 2012; Kang et al., 2016; Bae and Lee, 2017; Bae et al., 2021). Native speakers' mental lexicons are closely linked to SK morpheme understanding, regardless of Hanja literacy.[2] This raises an important question for AI: while LMs aim to mimic human cognition, do they truly process language like humans?

Studies on LM linguistic capabilities have largely focused on syntax and semantics, with limited attention to morpheme-level understand-

[1] Sino-Korean morphemes are Korean morphemes of Chinese origin, typically corresponding to Chinese characters, and they account for a large portion of Korean vocabulary.

[2] While SK morphemes can be represented via Hanja, we focus on Hangul text, as most Korean speakers acquire SK morphemes without Hanja exposure. Hanja is rarely used in modern Korean outside specialized contexts.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10762–10772
November 4-9, 2025

ing (Goldberg, 2019; Jawahar et al., 2019; Ettinger, 2020; Puccetti et al., 2021; Rogers et al., 2021). While these works explore syntactic dependencies and meaning composition, morphological structure—especially in morphologically rich languages like Korean—remains underexplored. SK morpheme comprehension requires syllable-level representation and morphological awareness, but naive syllable tokenization underperforms morpheme-based and BPE approaches in Korean NLP tasks (Park et al., 2020). Investigating whether LMs acquire such knowledge tests their alignment with human language acquisition and generalization beyond surface co-occurrence.

Building on these points, we ask whether Korean LMs utilize SK morphemes in language processing. Transformer-based LMs typically adopt Byte Pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2015) tokenizers. While full access to SK morphemes would require tokenizing each syllable, this is not feasible under BPE. Thus, it is worth examining whether LMs—despite lacking direct access to SK morphemes, which are essential for human-level understanding in Korean—still make use of them during processing.

For this purpose, we pose the research question: *Can a Transformer-based LM, without direct access to SK morphemes, still learn their compositional morphology?* We design experiments using a target SK word with a "real" neighbor (e.g., 생명 生命 *saengmyeong* 'life' & 생산 生産 *saengsan* 'production') that shares the same SK morpheme, 생 生 *saeng*, and a "fake" neighbor (e.g., 생명 生命 & 생략 省略 *saengnyak* 'omission') that shares only the Hangul syllable 생 *saeng*.

Using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), a classic Transformer model, alongside fastText (Bojanowski et al., 2017)—which indirectly accesses SK morphemes via Hangul syllables—yields mixed results. Both BERT and fastText assign higher similarity to real than fake pairs, suggesting that BERT may appear to handle the compositional morphology of SK morphemes even without explicit subword segmentation. However, further analysis reveals that BERT's performance is not due to genuine morphological understanding, but rather to its reliance on word frequency in the training corpus. While neither model truly understands SK morphemes, BERT's performance in particular is driven by exposure to frequent words, creating only the illusion of semantic understanding.

In this paper, we contribute the following:

- **Sino-Korean morphological approach in Transformer-based Korean models**: We provide an empirical investigation into how Transformer-based Korean LMs process the compositional morphology and semantics of SK morphemes, a topic that has received little attention in prior work.
- **Revealing the limits of purely distributional approach**: Our analysis shows that relying solely on distributional patterns is insufficient for learning compositional morphology, as these models struggle to capture SK morphemes.
- **Specialized dataset for Sino-Korean morphemes**: We present a small but focused dataset to evaluate LMs' compositional morphological knowledge, offering a starting point for targeted research on Korean morphology.

2 Related Work

2.1 Hanja-aware Approach for Korean Linguistic Tasks

Hanja-aware approaches aim to enhance semantic representation by supplementing Hangul input with corresponding Hanja forms—a key factor in SK morpheme understanding. Yoo et al. (2019) introduced a Hanja-aware fastText model that improved word-analogy and similarity tasks. Yoo et al. (2022) built HUE, a pretrained BERT model for Hanja texts, showing that Hanja-specific training aids historical document analysis. Yang et al. (2023) presented HistRed, a relation-extraction dataset with bilingual Hanja–Korean annotations, revealing that Hanja-based models surpass monolingual baselines in entity relation tasks.

Hanja-aware methods have also been explored in Neural Machine Translation (NMT). Kim et al. (2020) proposed a preprocessing step using Hangul-to-Hanja conversion to improve BLEU scores in Korean-to-Japanese translation. Son et al. (2022) introduced H2KE, a historical-to-modern Korean model that uses Hanja to better align older texts with modern usage.

These studies suggest that incorporating Hanja helps Korean LMs, but it is unclear whether they process SK morphemes like humans or just treat Hanja as extra vocabulary items.

2.2 Extracting Word Embedding Vectors from Contextualized Representations

While extracting word embeddings from static models (e.g., fastText) is straightforward, contextualized models like BERT produce token representations that vary with context. Mickus et al. (2020) used embeddings from the last layer of `bert-large-uncased` to examine semantic coherence in BERT's space. Studying five pretrained LMs, Bommasani et al. (2020) found that mean-pooling over 100K sentences yields more stable embeddings resembling static ones. Gupta and Jaggi (2021) also proposed a method to extract stable, static embeddings from contextualized models by pooling vectors across contexts containing each target word.

3 General Methodology

This study examines whether an LM can learn compositional morphology purely from word distributions in Hangul-only text. We construct word pairs to test whether a model can detect SK morphemes in cases of Hangul homography. The idea is simple: we compare "real" pairs (same syllable and SK morpheme) and "fake" pairs (same syllable, different morphemes). A model that captures SK morphemes should assign higher similarity to real pairs.

This section outlines shared methodological details across all experiments, including word pair selection (Section 3.1), model selection and tokenization strategy (Section 3.2), embedding extraction (Section 3.3), and Intrinsic Evaluation Approach (Section 3.4). Specific evaluation methods and metrics are discussed in Section 4, where each experiment is detailed separately.

3.1 Word Pair Selection

Focus on the First Syllable We focus on the first syllable for word pair selection. Psycholinguistic research, such as Cutler and Norris (1988), has consistently shown that initial syllables serve as crucial cues in lexical access. Similarly, Korean studies highlight the importance of the first syllable: (Kwon et al., 2011; Nam, 2022; Lee et al., 2023). Based on these findings, we restricted neighbor selection to words sharing the first syllable in order to ensure tighter experimental control.[3]

[3]While the present study focuses on first-syllable overlap, the position of the target syllable may affect the results. Future work could test whether final-syllable overlap yields different patterns.

Analysis Framework for Sino-Korean Words The dataset consists of two-syllable SK nouns, a common structure in Korean. As mentioned above, we selected words whose first syllable—an SK morpheme—governs the pairing. To prevent the model from relying on simple recognition rather than true SK morpheme understanding, we excluded words whose first syllable is both a fully independent morpheme and a standalone word—such as 왕 (王 *wang* 'king') and 외 (外 'outside').

To ensure linguistic validity, we consulted recent Korean linguistics studies on SK morphology. Since the linguistic treatment of SK morphemes vary among researchers,[4] we mainly focus on two recent PhD dissertations (Yang, 2010; He, 2018). Following the framework in He (2018), our experimental words cover three SK morpheme types:

- **Nominal Roots**: Dependent morphemes found in multi-syllable SK words, such as 천 (天 *cheon* 'sky') in 천국 (天國 *cheon-guk* 'heaven') and 인 (人 *in* 'person') in 인구 (人口 *in-gu* 'population').

- **Verbal/Adjectival Roots**: Bound stems used in verbs and adjectives, such as 강 (強 *gang* 'strong') in 강조 (強調 *gang-jo* 'emphasis') and 변 (變 *byeon* 'change') in 변신 (變身 *byeon-sin* 'transformation').

- **Prefixes**: Morphemes functioning as prefixes, such as 초- (超 *cho* 'super') in 초음속 (超音速 *cho-eum-sog* 'supersonic') and 신- (新 *sin* 'new'), 신세대 (新世代 *sin-se-dae* 'new generation').

Although SK morphemes have many subcategories, we do not separate them in our experiments since the classification remains debated in Korean linguistics as mentioned above. Accordingly, we treat them as dependent morphemes that cannot stand alone as words.

Neighbor Word Selection For each target word, we chose neighbor words based on the Hangul spelling of their first syllable, ensuring a controlled comparison of morphemic relationships. Only two-syllable SK nouns were considered, strictly matching the target's first syllable.

[4]For example, 최 (最 *choe* 'the most') is classified as a root in Yang (2010), but as a prefix in He (2018) and the *Standard Korean Language Dictionary*.

Figure 2: Examples of real and fake neighbor word pairs. The target word 신작 (新作 *sin-jag* 'new work') and its real neighbor 신입 (新入 *sin-ib* 'newcomer') share both the first Hangul syllable and the SK morpheme. In contrast, the fake neighbor 신용 (信用 *sin-yong* 'trust') share only the Hangul syllable but derive from different SK morphemes, making them unrelated in meaning.

We classified neighbor words into two types: "real" neighbors, which share the same SK morpheme (e.g., 신 작 新 作 *sin-jag* 'new work' & 신입 新入 *sin-ib* 'newcomer'), and "fake" neighbors, which share only the Hangul syllable but not the SK morpheme (e.g., 신 작 新 作 & 신 용 信 用 *sin-yong* 'trust'). Figure 2 illustrates this difference. Note that SK morphemes are represented by a single Hangul syllable, but one syllable can correspond to multiple distinct Hanja morphemes. Thus, words may share the same syllable without sharing the same morpheme, which is the basis of our real vs. fake distinction.

Final Word Pair Selection To ensure stable embeddings, we chose only high-frequency words from the training corpus. To reduce confounds from homography, we included words with either no homographs or only one highly frequent sense. Based on the Modern Korean Usage Frequency Survey (Kim, 2005), we chose words with at most two distinct senses in actual usage, ensuring minimal ambiguity while maintaining lexical coverage.

Most fake neighbor words' first syllables are SK morphemes. When no suitable match meets the criteria, we allow words whose first syllables are not SK morphemes. The final dataset consists of 100 target words, 100 real neighbors, and 100 fake neighbors, enabling a controlled evaluation. The neighbor list is fixed and independent of the model (BERT or fastText).

For simplicity, we call the real and fake pairs in BERT and fastText as *BERT-Real*, *BERT-Fake*, *fT-Real*, and *fT-Fake*.

3.2 Model Selection and Tokenization Strategy

Transformer-based model: BERT We use BERT as our Transformer-based model, chosen for its efficiency in small-scale experiments. While BERT typically uses a BPE tokenizer that may split SK words, we instead apply a morphological analyzer to ensure each SK word is tokenized as a single unit. This controlled setup allows us to test whether the model learns morphemic meaning from distributional patterns alone.[5] It also ensures a fair comparison with fastText, which produces the same segmentation but encodes subword-level information.

Specifically, we use the $\text{BERT}_{\text{BASE}}$ `Morpheme` model from Park et al. (2020), which uses MeCab-ko[6] for tokenization. For example, an SK word 생산하며 (生産하며 *saeng-san-ha-myeo* 'produce and') is tokenized as 생산, 하, 며, while the default BPE tokenizer splits it as 생산, ## 하며.

Baseline model: fastText As our baseline, we use fastText, which represents words via subword units. While it can directly access syllables, it lacks explicit SK morpheme segmentation unless Hangul-to-Hanja conversion is applied. This setup tests whether morphemic understanding arises purely from distribution or benefits from syllable-level access. Despite this, we expect BERT to outperform fastText due to its general superiority in NLP tasks.

We use the `morpheme_mecab_orig_composed` variant from Jeon (2022), which also employs MeCab-ko for tokenization. Its training corpus substantially overlaps with that of the BERT model (Park et al., 2020), ensuring comparability.

3.3 Embedding Extraction

We use example sentences from the *Standard Korean Language Dictionary*[7] to derive BERT embeddings for the experimental words. If no suitable example sentences are available, we supplement them with sentences from *Urimalsem*[8] and the *Korea University Korean Dictionary*, accessed via the Naver Korean Dictionary platform[9].

[5] Although the morpheme tokenization setting is somewhat idealized in terms of standard Transformer-based LMs, Park et al. (2020) reported minimal performance differences between morpheme and BPE tokenization.

[6] https://bitbucket.org/eunjeon/mecab-ko

[7] https://stdict.korean.go.kr/

[8] https://opendict.korean.go.kr/main

[9] https://ko.dict.naver.com/

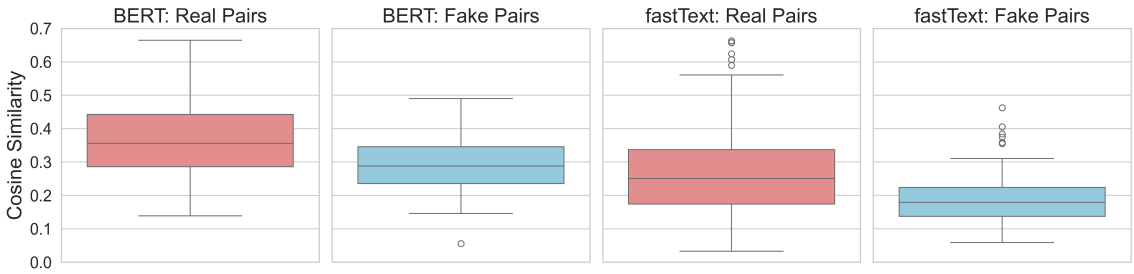

Figure 3: Cosine similarity distributions for real and fake pairs in BERT and fastText. Real pairs show higher similarity than fake pairs, suggesting that both models capture meaningful semantic relationships to a certain extent.

Specifically, we average each word's embeddings across its example sentences and concatenate the last four layers, following Devlin et al. (2019). For fastText, which is non-contextual, we extract static embeddings directly from its embedding set.

3.4 Intrinsic Evaluation Approach

We adopt an intrinsic evaluation strategy without any downstream classification or fine-tuning. Our aim is not to test whether the models can be trained to learn morphemic meanings, but whether they already acquire them solely through pretraining.

4 Experiments

To investigate SK morpheme understanding in Transformer-based LMs without direct access to morpheme syllables, we conduct three experiments. Experiment 1 (Section 4.1) tests whether the models distinguish real from fake word pairs via cosine similarity. Experiment 2 (Section 4.2) compares Top-K *similar words* for each pair to see if shared morphemes yield more overlap. Experiment 3 (Section 4.3) uses linear regression to test whether frequency—crucial in human language acquisition (Lieven, 2010; Ambridge et al., 2015)—also affects LMs.

4.1 Experiment 1: Cosine Similarity

We compute cosine similarity between word pairs to test whether models capture compositional morphology of SK morphemes. Each target word is compared to a real neighbor (e.g., 생명 生命 *saengmyeong* 'life' & 생산 生産 *saengsan* 'production') and a fake neighbor (e.g., 생명 & 생략 省略 *saengnyak* 'omission'). If real pairs consistently show higher similarity, this suggests the model encodes morphemic meaning beyond orthographic overlap.

Setup Recall from Section 3.1 that we use 100 target SK words, each paired with 100 real and 100 fake neighbors, yielding 100 similarity checks per model. We extract embeddings from BERT and fastText (Section 3.3), compute cosine similarity, and evaluate results in two ways: (1) *accuracy*, i.e., in a binary comparison, whether a real pair has higher similarity than a fake pair, and (2) statistical analysis, which compares cosine similarity distributions across real and fake pairs to determine significance. In (2), we use the Wilcoxon signed-rank test to compare cosine similarity for each group.

Results BERT scores 75% accuracy, fastText 71%, both above chance (50%), implying they might capture some SK morpheme meaning. BERT, despite lacking direct access to syllables and SK morphemes, slightly outperforms fastText. This suggests that explicit subword segmentation in fastText does not provide a significant advantage in distinguishing SK morphemes. However, given BERT's general superiority in NLP tasks, the difference is not substantial enough to be considered meaningful.

Figure 3 presents the cosine similarity distributions. Real pairs show significantly higher similarity than fake pairs for both BERT ($W = 907.0$, $p < .001$) and fastText ($W = 1100.0$, $p < .001$). In the model-wise comparison, BERT shows higher overall similarity than fastText ($W = 631.0$, $p < .001$ for real, $W = 399.0$, $p < .001$) for fake, indicating BERT generally assigns higher similarity between words.

While BERT shows stronger overall associations, this does not imply better SK morpheme understanding—it likely reflects BERT's general strength in lexical representation. The key question is whether these similarities indicate morphemic sensitivity, which we test through pairwise contrasts and other experiments.

4.2 Experiment 2: Similar Words Retrieval

To assess whether the models truly understand SK morphemes, Experiment 2 examines similar word retrieval: for each pair, we compute shared similar words—overlapping Top-K words by cosine similarity—and check how many share the target's first SK morpheme. This tests whether models implicitly group words by morphemic meaning without explicit clustering.

Setup We analyze *shared similar words* between real and fake pairs using BERT and fastText embeddings. For each pair, we retrieve the top $K = 600$[10] similar words from a 19,591-word SK noun vocabulary. *Shared similar words* are those appearing in both the target's and neighbor's lists. Among them, *SK-Morpheme Similar Words* start with the same Hanja syllable as the target. For example, for the real pair 생명 (生命 *saeng-myeong* 'life') and 생산 (生産 *saeng-san* 'production'), 48 similar words overlap, and 4 of those share the SK morpheme 생 (生 *saeng*), such as 생활 (生活, *saeng-hwal* 'living').

This allows us to measure:

1. The total number of *shared similar words* between the target and neighbor.
2. The proportion of *SK-Morpheme Similar Words* among all *shared similar words*.

For the *pair-wise* comparison (real vs. fake), we use the Wilcoxon signed-rank test to check whether real pairs yield more *shared similar words*. To reduce noise from pairs with very few shared similar words, we exclude those below the 25th percentile. For the *model-wise* comparison (BERT vs. fastText), we again use the Wilcoxon test on both *shared similar words* and *SK-Morpheme Similar Words*. After filtering, 46 real and 71 fake pairs remain.

Results Table 1 shows the pair-wise comparison: both *BERT-Real* and *fT-Real* retrieved significantly more *shared similar words* than their fake counterparts ($p < .001$), suggesting that semantically real pairs tend to share more neighbors. However, Table 2 shows no significant difference in the proportion of *SK-Morpheme Similar Words* between real and fake pairs; thus, we cannot conclude that the models reliably capture SK morpheme information reflected in pair types.

[10] Ablation analysis determined this value for balanced sample size and semantic quality.

	Mean (SD)	Median (IQR)	p-value
Pair-wise			
BERT-Real	113.50 (70.40)	98.50 (92.75)	$< .001$
BERT-Fake	40.57 (43.28)	23.00 (31.50)	
fT-Real	92.03 (84.94)	60.50 (87.00)	$< .001$
fT-Fake	30.21 (31.19)	16.50 (36.00)	
Model-wise			
Real-BERT	105.35 (64.36)	89.50 (90.50)	$< .05$
Real-fT	87.15 (82.14)	59.00 (75.50)	
Fake-BERT	78.14 (65.44)	54.00 (83.50)	.058
Fake-fT	66.99 (75.79)	39.00 (73.50)	

Table 1: Comparison of the number of *shared similar words* for real and fake pairs in BERT and fastText. SD: standard deviation; IQR: interquartile range.

	Mean (SD)	Median (IQR)	p-value
Pair-wise			
BERT-Real	2.74 (2.54)	2.55 (3.75)	.082
BERT-Fake	1.53 (4.06)	0.00 (0.00)	
fT-Real	2.82 (3.31)	1.30 (4.49)	.059
fT-Fake	2.62 (6.40)	0.00 (1.35)	
Model-wise			
Real-BERT	2.66 (3.01)	2.03 (3.94)	.323
Real-fT	3.07 (3.56)	1.55 (4.89)	
Fake-BERT	2.12 (2.88)	0.63 (3.27)	.142
Fake-fT	3.56 (6.86)	0.47 (4.47)	

Table 2: Comparison of the proportion (%) of *SK-Morpheme Similar Words* among *shared similar words* in BERT and fastText. SD and IQR as in Table 1.

Table 1 also presents the model-wise comparison. *BERT-Real* retrieves significantly more *shared similar words* than *fT-Real* ($p < .05$), highlighting BERT's strength in capturing lexical-semantic relationships. In Table 2, *fT-Fake* shows a higher proportion of *SK-Morpheme Similar Words* than *BERT-Fake* ($p < .05$). Although this may suggest stronger morphological representation, it actually reflects fastText's subword segmentation capturing homography rather than true SK morpheme understanding. There is no linguistic reason for a higher proportion in fake pairs.

Overall, the results show that BERT retrieves more *shared similar words* than fastText, highlighting its strength in capturing lexical-semantic relationships. However, across all comparisons, BERT shows no clear advantage in capturing *SK morpheme* meaning. Thus, although Experiment 1 showed above-chance accuracy, neither model systematically distinguishes SK morphemes in retrieved *similar words*, so this accuracy should not be taken as true *SK morpheme* comprehension.

	BERT (Real Pair)	*BERT (Fake Pair)*	*fastText (Real Pair)*	*fastText (Fake Pair)*
Intercept	0.352 ($p < .001$)	0.271 ($p < .001$)	0.367 ($p < .001$)	0.204 ($p < .001$)
`Target Frequency`	**14310** ($p < .01$)	6325 ($p = .241$)	4144 ($p = .483$)	-137 ($p = .981$)
`Neighbor Frequency`	314 ($p = .749$)	-480 ($p = .461$)	441 ($p = .687$)	49 ($p = .945$)
`SK Neighbor Frequency`	-53.43 ($p = .152$)	3.16 ($p = .906$)	**-193.36** ($p < .001$)	-33.94 ($p = .256$)
`Bound morpheme + Free morpheme`	-0.016 ($p = .615$)	**0.057** ($p < .05$)	**-0.098** ($p < .01$)	-0.012 ($p = .633$)
R^2	0.203	0.149	0.398	0.031

Table 3: Regression results for cosine similarities in BERT and fastText embeddings. Significant predictors ($p < .05$) are bolded.

4.3 Experiment 3: Regression Analysis

In Experiment 3, we apply Ordinary Least Squares (OLS) regression to examine whether the same factors known to guide human learning of SK morphemes also shape how language models represent and process them.

Hypothesis Based on psycholinguistic findings that repeated exposure leads to more robust lexical representations in humans(Lieven, 2010; Ambridge et al., 2015), we propose the following hypotheses:

- **Hypothesis 1**: Higher frequencies of both a target word and its neighbors lead to more stable embeddings and thus stronger similarity.[11]
- **Hypothesis 2**: This frequency effect holds primarily for real pairs, not for fake pairs, which lack meaningful overlap.

Setup We converted raw word frequencies into relative frequencies, applied a $\log(1 + x)$ transformation, and removed outliers using the IQR method—excluding values outside $(Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR})$. This filters out overly frequent or infrequent words—whose embeddings may be overly stable or unreliable—yielding 56 real and 52 fake pairs.

SK neighbors are nouns or roots sharing the target's first SK morpheme (e.g., 생활 生活 *saenghwal* 'living' or 생물체 生物體 *saengmulche* 'organism' for 생명 生命 *saengmyeong* 'life'). Using the *Standard Korean Language Dictionary* and Kim (2005), we found words matching the target's first SK syllable and counted their corpus occurrences. Since the corpus lacks explicit Hanja characters, these frequency counts are approximate and may include homographs.

[11] Although cosine similarity normalizes magnitude, higher frequency is expected to yield more stable embeddings by reinforcing co-occurrence patterns.

We analyze real and fake pairs with separate regression models, assuming frequency raises similarity for real pairs but not for fake ones. Thus, we run four models: *BERT (Real)*, *BERT (Fake)*, *fastText (Real)*, and *fastText (Fake)*.

Key Variables

1. **Continuous** `Target Frequency`: Frequency of the target word in the training corpus
2. **Continuous** `Neighbor Frequency`: Frequency of the neighbor word in the training corpus
3. **Continuous** `SK Neighbor Frequency`: Total token frequency of all SK neighbors of the target word in the training corpus
4. **Categorical** Target word's SK morpheme structure
   - `Bound morpheme + Bound morpheme` (e.g., 국립 國立 *gug-lib* 'national') → Coded as 0
   - `Bound morpheme + Free morpheme` (e.g., 등산 登山 *deung-san* 'mountain climbing') → Coded as 1

The categorical variable is introduced to assess differences in syllable accessibility. In fastText, free morphemes can function as standalone tokens, allowing direct access. If fastText shows performance changes while BERT does not, this implies that direct syllable-level access influences SK morpheme comprehension—even without explicit Hanja. By contrast, BERT lacks this direct access and may struggle to capture morphemic meaning.

Results Table 3 summarizes the findings. In *BERT (Real Pair)*, `Target Frequency` has a significant positive effect on cosine similarity ($\beta = 14310$, $p < .01$), implying that more frequent

words generally yield higher similarity scores. No other predictors are significant.

For *BERT (Fake Pair)*, the categorical variable `Bound morpheme + Free morpheme` shows a significant effect ($\beta = 0.057$, $p < .05$), but the low R^2 indicates weak explanatory power.

In *fastText (Real Pair)*, `SK Neighbor Frequency` exerts a strong negative effect on similarity ($\beta = -193.36$, $p < .001$), and `Bound morpheme + Free morpheme` also shows a significant negative effect ($\beta = -0.098$, $p < .01$). This suggests that free morphemes, which may appear as independent tokens, introduce noise that lowers cosine similarity.

For *fastText (Fake Pair)*, no predictor is significant and the model's $R^2 = 0.031$ is very low, suggesting that neither frequency nor morphology affects similarity in fake pairs. This contradicts the expectation that higher frequency lowers similarity for unrelated words, indicating both models show limited SK morpheme understanding.

Overall, BERT's similarity for real pairs hinges on frequency, highlighting its dependence on co-occurrence. In fastText, the categorical variable matters only in real pairs, where free morphemes reduce similarity—likely due to their separate-token status injecting noise into embeddings.

5 Discussion

Although both BERT and fastText show relatively high accuracy overall (Section 4.1), a closer look reveals that token frequency drives much of their performance, rather than a genuine grasp of SK morphemes. BERT, in particular, assigns higher similarity to frequently seen words, but this reflects context-driven process that more strongly associates words appearing together often than true morphemic awareness.

Greater exposure to a word improves its contextual representation, but does not directly reveal its internal compositional morphology. More importantly, the models do not appear to learn this structure even indirectly: they process each word's meaning in isolation rather than forming semantic clusters based on SK morpheme understanding.

A central issue is that language models cannot explicitly encode the shared structures linking multiple words with the same SK morpheme. Humans readily recognize that 국어 (國語 *gug-eo* 'national language'), 국립 (國立 *gug-lib* 'national'), and 국사 (國史 *gug-sa* 'national history') share the morpheme 국 (國 *gug* 'nation') phonologically and orthographically, whereas LMs fail to systematically detect such patterns. Unlike humans, Transformer-based models rely solely on contextual associations, overlooking information accessible through morphemic awareness.

This limitation affects not only our morpheme-tokenized model but also typical BPE-based models. Whereas humans refine their morphemic knowledge by repeatedly seeing words that share a syllable, LMs trained solely on co-occurrence data struggle to internalize morphological structure. Even large amounts of training data do not ensure a systematic grasp of morpheme-level structures, underscoring how purely distributional learning is insufficient for true understanding of compositional morphology of SK morphemes.

To overcome these limitations, explicitly encoding shared morphemic information is key to improving LMs'grasp of SK morphemes. Current tokenization methods, such as BPE, rely on statistical segmentation rather than morphological principles, limiting their ability to capture systematic SK relationships. While subword models (e.g., fastText) do include syllable-level access, our findings show that simply splitting by syllables is insufficient because one Hangul syllable may map to multiple distinct Hanja morphemes. Moreover, Park et al. (2020) demonstrate that syllable-level tokenization underperforms both morpheme-based and BPE approaches in many tasks.

A more effective approach may involve linguistically informed tokenization. For example, Kim et al. (2024) propose a phonology-aware method that separates Hangul onset-nucleus and coda, leading to improved performance over BPE-based models. This highlights the need for alternative tokenization strategies that explicitly integrate SK morphemes rather than relying solely on character sequences.

Considering the above, incorporating Hanja-aware models could enhance the performance of LMs. Research on human reading proficiency shows that less-proficient readers rely more on orthography, whereas more-proficient readers rely more on morphemes—likely using them to infer the meanings of unfamiliar words (Yi, 2009). Similarly, explicitly integrating Hanja information could help LMs capture deeper morphemic structures. Unlike human learners, who may find learning Hanja challenging, LMs face no such difficulty, making Hanja incorporation a promising approach

for advancing SK morpheme comprehension.

More broadly, although we test SK words in Korean, the issue is not language-dependent: it asks whether LMs relying solely on distributional information—often effective without true morphological understanding—can be said to process language as humans do. This question is not confined to Korean. Japanese also has numerous Sino-Japanese compounds written in Kanji (Japanese Chinese characters), with added complexity due to multiple readings. Even in English, BPE tokenizers sometimes cut across morpheme boundaries (e.g., "unbelievably → un + bel + ievably" instead of the human segmentation "un + believe + able + ly"). In such cases, the model faces a challenge similar to that of a Korean LM lacking understanding of SK morphemes. This underscores that the limitations we observed reflect a general property of distributional learning rather than an issue confined to Korean NLP.

Finally, while our study restricted neighbors to those sharing the first syllable, future work should test whether overlap in other positions (e.g., final syllables) yields different patterns since the position of the target syllable may affect the results. In addition, although most fake pairs are contextually unrelated, some may co-occur in similar domains (e.g., 대결 對決 *dae-gyeol* 'match' and 대회 大會 *dae-hoe* 'competition'), potentially inflating similarity scores.[12] Humans can readily distinguish these two types of similarity, contextual similarity vs. context-free similarity, but language models may conflate them; future studies should therefore compare model predictions with human judgments more directly.

6 Conclusion

While modern LMs successfully validate the core idea behind the distributional hypothesis (Firth, 1957; Sahlgren, 2006; Kornai, 2023; Jurafsky and Martin, 2024), distribution alone does not fully explain word meaning. Sino-Korean morphemes in Korean illustrate this gap. BERT, a monumental Transformer LM, does not seem to truly grasp their compositional morphology; rather, it learns each word separately, driven by frequency in training data. Achieving deeper, human-like language understanding thus requires more than distributional cues.

Limitations

This study has several limitations, primarily related to the size of our dataset, the availability of example sentences, and the absence of comparison with a Hanja-trained Transformer-based model.

First, the dataset is relatively small by NLP standards, although it may be sufficient for human subject experiments. This limitation arises from inconsistencies in how Korean linguistics researchers classify SK morphemes, particularly in distinguishing roots from affixes. To ensure consistency, we relied on previously illustrated examples in recent PhD dissertations on SK morphology (Yang, 2010; He, 2018), which led to a reduced dataset size. Additionally, since this study uses Hangul-only text and models, we had to limit experimental words to those with minimal homographic ambiguity or none at all to control for confounding factors. This further constrained the dataset.

Second, the number of example sentences for extracting embedding vectors is limited. On average, each target word has 4.2 sentences, and each word pair has 4.77 sentences, primarily sourced from dictionary examples. However, no existing corpus provides both Hanja-converted text and dictionary headword information, limiting access to richer contexts. Furthermore, some extracted examples had to be removed due to unreliable automatic morphological analysis, which failed to tokenize experimental words as single tokens in some cases. This further reduced the available data.

Finally, this study does not include a comparison with a Hanja-trained Transformer-based model, which could provide direct insight into whether explicit morpheme representations improve performance. To the best of our knowledge, no large-scale Hanja-Hangul-converted corpus exists for training such models, nor is there a Transformer-based LM specifically designed for Hanja-aware Korean text. Future research should explore constructing such models to directly assess the role of Hanja information in SK morpheme comprehension.

[12] We thank a reviewer for this insightful comment. Upon re-checking our fake pairs, we found that some, like 대결 (對決) and 대회 (大會), may indeed co-occur in similar contexts. However, we also confirmed that the vast majority of fake pairs in our dataset do not appear in comparable domains, so we believe this issue does not critically affect our results.

References

Ben Ambridge, Evan Kidd, Caroline F Rowland, and Anna L Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of child language*, 42(2):239–273.

Sungbong Bae and Donghoon Lee. 2017. Individual differences in the morphological decomposition of hanja words. *Korean Journal of Cognitive and Biological Psychology*, 29(4):455–462.

Sungbong Bae, Hye K Pae, and Kwangoh Yi. 2021. Modeling morphological processing in korean: Within-and cross-scriptal priming effects on the recognition of sino-korean compound words. *Reading and Writing*, pages 1–30.

Sungbong Bae, Kwangoh Yi, and HyeWon Park. 2012. Semantic transparency effects in the recognition and learning of sino-korean words. *Educational Psychology Research*, 26(2):607–620.

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.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Miho Choo and William O'Grady. 1996. *Handbook of Korean vocabulary: A resource for word recognition and comprehension*. University of Hawaii Press.

Anne Cutler and Dennis Norris. 1988. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human perception and performance*, 14(1):113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina, and Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

JR Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. *arXiv preprint arXiv:2106.04302*.

Chengjin He. 2018. *A Study on the Structures of Sino-Korean Word and Sino-Hybrid Word*. Ph.D. thesis, Seoul National University.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Taehee Jeon. 2022. A linguistic study on tokenization methods for korean text. *Language Facts and Perspectives*, 55:309–354.

Daniel Jurafsky and James H. Martin. 2024. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models*, 3rd edition.

Jinwon Kang, Sooleen Nam, Heui Seok Lim, and Kichun Nam. 2016. Erp indices of korean derivational prefix morphemes separated from the semantic and orthographic information. *The Korean Journal of Cognitive and Biological Psychology*, 28(3):409–430.

Hansaem Kim. 2005. *Modern Korean Usage Frequency Survey*. National Institute of Korean Language, Seoul.

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.

SungHo Kim, Juhyeong Park, Yeachan Kim, and SangKeun Lee. 2024. Kombo: Korean character representations based on the combination rules of subcharacters. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5102–5119.

András Kornai. 2023. *Vector semantics*. Springer Nature.

Youan Kwon, Yoonhyoung Lee, and Kichun Nam. 2011. The different p200 effects of phonological and orthographic syllable frequency in visual word recognition in korean. *Neuroscience letters*, 501(2):117–121.

Solbin Lee, Eun-Ha Lee, Joonwoo Kim, Sangyub Kim, Jeahong Kim, Jinwon Kang, Changhwan Lee, and Kichun Nam. 2023. The effect of the first syllable and syllables in other positions in visual word recognition of korean noun eojeol: Focusing on token frequency. *The Korean Journal of Cognitive and Biological Psychology*, 35(3):151–164.

Elena Lieven. 2010. Input and first language acquisition: Evaluating the role of frequency. *Lingua*, 120(11):2546–2556.

Timothee Mickus, Denis Paperno, Matthieu Constant, and Kees van Deemter. 2020. What do you mean, bert? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290.

Ki-Chun Nam. 2022. The first syllable frequency effect in korean morphologically complex word recognition associated with hemispheric dominance and coordination. *Journal of the Korea Academia-Industrial Cooperation Society*, 23(4):505–515.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various korean nlp tasks. *arXiv preprint arXiv:2010.02534*.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How do bert embeddings organize linguistic knowledge? In *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pages 48–57.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis of represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *Ph. D. thesis, Stockholm University*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary korean and english. *arXiv preprint arXiv:2205.10019*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. Histred: A historical document-level relation extraction dataset. *arXiv preprint arXiv:2307.04285*.

Yeong Jin Yang. 2010. *A Study on the Bound Morpheme of Sino-Korean Complex Words*. Ph.D. thesis, Gyeongsang National University.

Kwangoh Yi. 2009. 30 morphological representation and processing of sino-korean words. *The handbook of East Asian psycholinguistics*, page 398.

Kwangoh Yi, Jingab Jung, and Sungbong Bae. 2007. Writing system and visual word recognition: Morphological representation and processing in korean. *The Korean Journal of Experimental Psychology*, 19(4):313–327.

Kwangoh Yi and Insun Yi. 1999. Morphological processing in korean word recognition. *Korean Journal of Experimental and Cognitive Psychology*, 11(1):77–91.

Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Hue: Pretrained model and dataset for understanding hanja documents of ancient korea. *arXiv preprint arXiv:2210.05112*.

Kang Min Yoo, Taeuk Kim, and Sang-goo Lee. 2019. Don't just scratch the surface: Enhancing word representations for korean with hanja. *arXiv preprint arXiv:1908.09282*.