PhoniTale: Phonologically Grounded Mnemonic Generation for Typologically Distant Language Pairs

Sana Kang[1]* Myeongseok Gwon[1]* Su Young Kwon[1]*
Jaewook Lee[2] Andrew Lan[2] Bhiksha Raj[3] Rita Singh[3]
[1]KAIST [2]University of Massachusetts Amherst [3]Carnegie Mellon University
{sanakang0615, myeongseok, suyoungkwon}@kaist.ac.kr,
{jaewooklee, andrewlan}@cs.umass.edu, {bhiksha, rsingh}@cs.cmu.edu

Abstract

Vocabulary acquisition poses a significant challenge for second-language (L2) learners, especially when learning typologically distant languages such as English and Korean, where phonological and structural mismatches complicate vocabulary learning. Recently, large language models (LLMs) have been used to generate keyword mnemonics by leveraging similar keywords from a learner's first language (L1) to aid in acquiring L2 vocabulary. However, most methods still rely on direct IPA-based phonetic matching or employ LLMs without phonological guidance. In this paper, we present PhoniTale, a novel cross-lingual mnemonic generation system that performs IPA-based phonological adaptation and syllable-aware alignment to retrieve L1 keyword sequence and uses LLMs to generate verbal cues. We evaluate PhoniTale through automated metrics and a short-term recall test with human participants, comparing its output to human-written and prior automated mnemonics. Our findings show that PhoniTale consistently outperforms previous automated approaches and achieves quality comparable to human-written mnemonics.

1 Introduction

Vocabulary acquisition remains one of the most persistent challenges for second-language (L2) learners. A classic—and surprisingly durable—strategy is *keyword mnemonic*: learners associate a new L2 lexical item with a familiar first-language (L1) word or phrase whose pronunciation is similar, and then build a vivid verbal or visual scene that links the two (Atkinson and Raugh, 1975). For example, a German learner might associate the word *Flasche* (bottle) with the phonetically similar English word *flashy*, forming the mnemonic *a flashy bottle that stands out from the rest*. This technique leverages phonological similarity while establishing a memorable semantic connection between the L2 target and L1 knowledge (Lee and Lan, 2023).

Typically, the L1 phrase corresponding to a given L2 term to be memorized is manually designed; however, this is a laborious process that scales poorly, necessitating the development of automated mechanisms to compose these phrases. Methods for automated generation of such keywords began with TransPhoner, which leverages the International Phonetic Alphabet (IPA) (International Phonetic Association, 1949) and hand-crafted heuristics to retrieve pronunciation-similar L1 keyword for target English words, resulting in significant recall gains (Savva et al., 2014).

Leveraging these methods, recent studies employ large language models (LLMs) for automated keyword generation. SmartPhone fed the TransPhoner keyword into GPT-3 to automatically generate verbal cues and DALL·E to generate visual cues (Lee and Lan, 2023). Lee et al. (2024) introduced an overgenerate-and-rank approach, where LLMs overgenerate keyword sequences and verbal cues, and then rank them according to multiple different criteria. Balepur et al. (2024) aligned mnemonics with user preferences by fine-tuning Llama 2 for personalization and cost-efficiency. Lee et al. (2025) learned latent user and Kanji (Chinese characters in Japanese) traits from a crowd-sourced platform for learning Kanji, and extract rules for constructing mnemonics using an Expectation-Maximization style algorithm.

These prior work focus predominantly on Indo-European L1-L2 language pairs with substantial phonological overlap (Savva et al., 2014; Lee and Lan, 2023). However, typologically distant language pairs, such as English-Korean, present unique challenges that remain underexplored. English and Korean exhibit four major phonological mismatches that make mnemonic generation challenging. First, orthographic systems differ in dimensionality because English prints letters lin-

*These authors contributed equally to this work.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25561–25593
November 4-9, 2025 ©2025 Association for Computational Linguistics

early, while Korean arranges the jamo into two-dimensional syllable blocks (Park and Li, 2009). Second, Korean forbids consonant clusters within a syllable, so epenthetic vowels must be inserted when adapting cluster-rich English words, which expands the syllable count (Kang, 2003; Kenstowicz, 2005). Third, certain English phonemes such as /θ/ have no direct Korean counterpart and are usually replaced with /s/ or /t/ (Tak, 2012; Kim and Kochetov, 2011). Fourth, the two languages exhibit phonemic contrast differences: Korean employs a three-way lenis, fortis, and aspirated stop contrast, whereas English distinguishes only voiced versus voiceless stops, and treats aspiration as a position-dependent allophone (Kang, 2014; Kang et al., 2022). These differences complicate the generation of phonologically faithful keyword sequence when Korean speakers learn English.

Contributions In this paper, we introduce a mnemonic generation system for language learning, PHONITALE. Our approach employs a greedy search through phonetically and syllabically approximated L2 sequences to identify the most suitable L1 keyword sequence. Specifically, we first transliterate L2 phonemes into L1-adapted sequences, segment these into syllables according to L1 phonological constraints, and then select keywords that maximize phonetic similarity while preserving syllabic structure. Unlike previous approaches that rely heavily on LLMs for keyword generation, we utilize LLMs only for verbal cue generation, while our specialized modules handle the cross-lingual phonological alignment. This design addresses the unique challenges posed by typologically distant languages, improves scalability, and mitigates hallucination risk. Through systematic evaluation including both automated metrics and human studies with short-term recall tests, we demonstrate that PHONITALE achieves comparable performance to human-authored mnemonics.

2 Problem Statement

PHONITALE performs the task of retrieving cross-lingual phonologically similar keyword sequence and using them to construct a L1 verbal cue for a given L2 target word, following the process illustrated in Figure 1. Let $w_{\text{L2}} \in \mathcal{V}_{\text{L2}}$ be a word in the L2, and let ℓ denote its meaning. The goal is to retrieve a L1 keyword sequence $\mathcal{W}_{\text{L1}} = (w_1^*, w_2^*, \dots) \in \mathcal{V}_{\text{L1}}$ that are phonologically similar to segments of w_{L2}, and to use $\mathcal{W}_{\text{L1}}$ to construct a verbal cue $c^* \in \mathcal{C}$, where $\mathcal{C}$ is the space of natural-language expressions in L1.

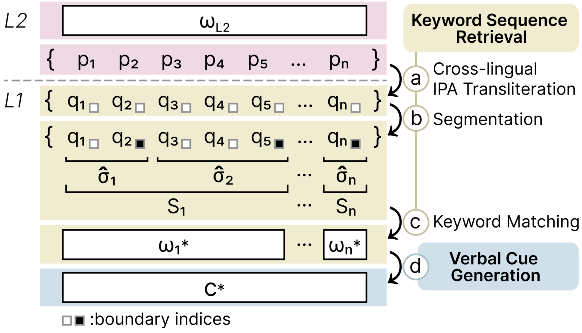


Figure 1: Problem formulation of the PHONITALE system. Phase 1, keyword sequence retrieval, comprises **(a)** IPA transliteration, **(b)** segmentation, and **(c)** keyword matching. Phase 2, **(d)**, performs verbal cue generation.

Retrieving phonologically similar keyword sequence begins by extracting the phoneme sequence of the L2 word, denoted $P_{\text{L2}} = (p_1, p_2, \dots, p_m)$, where each $p_j \in \Sigma_{\text{L2}}$, the L2 phoneme inventory. This sequence is then transliterated into an L1-adapted phoneme sequence $\widehat{P}_{\text{L1}} = (q_1, q_2, \dots, q_n)$, with $q_i \in \Sigma_{\text{L1}}$ as shown in Figure 1a, to approximate the L2 pronunciation using L1 phonological constraints.

Next, the adapted sequence is syllabified according to L1 phonological constraints. The syllabification process is non-deterministic, with multiple possible ways to divide the phoneme sequence. Each division creates different phoneme graphs that represent potential syllabification paths. From these multiple possibilities, a single path is selected that most closely aligns with L1 phonological patterns. This process yields the syllable sequence $\widehat{\boldsymbol{\sigma}} = (\sigma_1, \sigma_2, \dots, \sigma_l)$.

These syllables are then grouped into k segments $S_1, S_2, \dots, S_k$ using predefined partitioning rules. Each segment S_i consists of one or more complete syllables and is defined by boundary indices $0 = b_0 < b_1 < \cdots < b_k = l$ such that $S_i = (\sigma_{b_{i-1}+1}, \sigma_{b_{i-1}+2}, \dots, \sigma_{b_i})$. The complete segmentation process is represented in Figure 1b.

Subsequently, each segment S_i is then mapped to a keyword $w_i^* \in \mathcal{V}_{\text{L1}}$ whose pronunciation closely resembles S_i according to our phonological similarity criterion, as illustrated in Figure 1c. Finally, the verbal cue c^* is generated by embedding the keyword sequence $\mathcal{W}_{\text{L1}}$ in a natural-language expression that helps the learner associate the form of the L2 word with its meaning ℓ, completing the process shown in Figure 1d. The complete output is the pair $(\mathcal{W}_{\text{L1}}, c^*)$, which together support recall of the L2 word through phonological association.

3 Methodology

We divide the task into two phases: first, retrieving $\mathcal{W}_{\text{L1}}$ for a given w_{L2} using our keyword sequence retrieval component; and second, using LLMs to generate verbal cues from $\mathcal{W}_{\text{L1}}$, selecting the most coherent cue based on a ranking criterion.

3.1 Keyword Sequence Retrieval

We implement three modules by following the three steps of transforming w_{L2} into $\mathcal{W}_{\text{L1}}$.

3.1.1 Cross-lingual IPA Transliteration

In the IPA transliteration module, we convert L2 phoneme sequence P_{L2} of w_{L2} into their L1-adapted sequence P_{L1}. We utilize a neural sequence-to-sequence architecture with attention for the transduction task. We employ a bidirectional LSTM encoder and a unidirectional LSTM decoder, each with 256 hidden units (Bahdanau et al., 2014). The encoder processes P_{L2}, capturing contextual information from both directions to produce a 512-dimensional representation, which the decoder uses to generate $\widehat{P}_{\text{L1}}$.

We train the module by combining cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and contrastive loss ($\mathcal{L}_{\text{cont}}$):

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathbf{x}) \tag{2}$$

$$\mathcal{L}_{\text{cont}} = 1 - \frac{\mathbf{z}_{\text{enc}} \cdot \mathbf{z}_{\text{dec}}}{\|\mathbf{z}_{\text{enc}}\|_2 \cdot \|\mathbf{z}_{\text{dec}}\|_2} \tag{3}$$

The $\mathcal{L}_{\text{CE}}$ ensures token-level generation accuracy, while the $\mathcal{L}_{\text{cont}}$ promotes phonological consistency by aligning the encoder and decoder representations, $\mathbf{z}_{\text{enc}}$ and $\mathbf{z}_{\text{dec}}$, which are obtained by projecting their outputs into a shared embedding space via a lightweight feedforward layer. We combine these losses as a weighted sum, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{cont}}$, where we set $\lambda = 0.1$. We determined this weighting coefficient by analyzing the transliteration quality between $\widehat{P}_{\text{L1}}$ and ground-truth L1-adapted phoneme sequence P_{L1} on our development set.

3.1.2 Segmentation

In our segmentation module, we predict k contiguous phoneme segments S from $\widehat{P}_{\text{L1}}$ through a two-stage process. The first stage segments $\widehat{P}_{\text{L1}}$ into a syllable sequence $\widehat{\boldsymbol{\sigma}} = (\sigma_1, \sigma_2, \ldots, \sigma_l)$, where each $\sigma_i \in \Sigma_{\text{L1}}^{+}$ constitutes a valid L1 syllable. We assign binary labels to token positions in P_{L1} to indicate syllable boundaries, utilizing a bidirectional LSTM network with 256 hidden units in each direction. The network processes embedded IPA tokens augmented with binary vowel masks that signal potential syllabic nuclei (Mayer and Nelson, 2020).

The second stage combines these syllables into at most two segments (S_1, S_2) for lengthy w_{L2} words, addressing the fundamental challenge that one-to-one mapping with $\mathcal{W}_{\text{L1}}$ proves ineffective due to phoneme combinations or syllable structures in L2 that are absent in L1 (Daland and Zuraw, 2013). Each segment S_i represents a contiguous sequence of L1-adapted syllables derived from the original L2 word. For instance, L1-adapted phoneme sequence derived from the English word *autopsy* might be syllabified as /o/ - /t$^{\text{h}}$ɑp/ - /si/ and subsequently segmented into two segments such as S_1=/ot$^{\text{h}}$ɑp/ and S_2=/si/, or alternatively S_1=/o/ and S_2=/t$^{\text{h}}$ɑpsi/. This binary constraint prevents excessive fragmentation while preserving phonetic similarity and phonological coherence. The resulting segment sequence serves as input to the subsequent keyword sequence retrieval module, facilitating phonologically informed matching between w_{L2} and $\mathcal{W}_{\text{L1}}$.

3.1.3 Keyword Matching

In the keyword matching module, we calculate phonological similarity between each segment S_i and potential $\mathcal{W}_{\text{L1}}$ from $\mathcal{V}_{\text{L1}}$ to identify the most suitable matches. We convert phoneme sequences into 22-dimensional phonological feature vectors using PanPhon (Mortensen et al., 2016), capturing distinctive phonological characteristics. The similarity between a segment S_i and a candidate keyword $\mathcal{W}_{\text{L1},i}$ from Korean dictionary dataset (Ha, 2023) is computed using the cosine similarity of their phonological feature embeddings with a structural alignment adjustment:

$$\phi(S_i, \mathcal{W}_{\text{L1},i}) = \cos\left(\mathbf{v}(S_i), \mathbf{v}(\mathcal{W}_{\text{L1},i})\right) + \Delta_{\text{structural}} \tag{4}$$

where $\mathbf{v}(\cdot)$ represents the phonological feature embedding function and $\Delta_{\text{structural}}$ provides structural alignment adjustments.

While cosine similarity captures general phonological resemblance, this metric fails to account for syllable-level perception critical to Korean speakers (Siew et al., 2021; Lee and Taft, 2017; Yoon and Bolger, 2015; Kang, 2003). We incorporate four structural alignment adjustments in $\Delta_{\text{structural}}$: syllable overlap, initial-syllable match, early-phone alignment, and substring inclusion. Korean speakers perceive words as syllable bundles rather than phoneme strings, necessitating these adjustments to align our similarity function with native phonological perception processes. The initial-syllable match receives the highest weighting due to its greater perceptual significance in word recognition (Lee and Taft, 2017).

Using the similarity function ϕ, we pick the best keyword $\mathcal{W}_{\mathrm{L1},i}$ for each segment S_i by

$$\mathcal{W}^{*}_{\mathrm{L1},i} = \arg\max_{w \in \mathcal{V}_{\mathrm{L1}}} \phi(S_i, w), \tag{5}$$

and score an entire segmentation by averaging each segment's top match:

$$\frac{1}{m} \sum_{i=1}^{m} \max_{k \in \mathcal{V}_{\mathrm{L1}}} \mathrm{sim}(S_i, \mathcal{W}_{\mathrm{L1},i}). \tag{6}$$

This ranking process identifies the best keyword sequence $\mathcal{W}_{\mathrm{L1}} = (w_1^*, w_2^*)$ from our predefined segmentation candidates, selecting the keyword sequence that maximizes phonological similarity between the L2 word and the L1 keyword sequence.

3.2 Verbal Cue Generation

The verbal cue generation component builds upon Lee et al. (2024), while introducing methodological refinements specific to the Korean-English language pair. We implement two major modifications to adapt the approach to a cross-lingual setting.

First, in the prompt, we eliminate the two-step approach used in Lee et al. (2024) which first generates a story and then summarizes that story to produce a verbal cue. While this approach aims to preserve keyword sequence in complex verbal cues, our cross-lingual setting with only two keywords, making this constraint unnecessary. We therefore directly generating without summarization which we validate through ablation studies presented in Section 5.4 (Appendix Table 5 for the prompt).

Second, we discard the Age-of-Acquisition (AoA) ranking criterion from Lee et al. (2024) as it does not generalize effectively to cross-lingual contexts. The AoA of a word in L2 fails to reliably reflect its familiarity in L1. We retain only the context completeness criterion, calculating this by masking the target word in the verbal cue and prompting GPT-4o (OpenAI, 2024) to generate five probable candidates. We then compute the average cosine similarity between FastText (Bojanowski et al., 2016) embeddings of these candidates and the target word, trained on Korean corpus data (Lee, 2020). This approach quantifies how effectively the verbal cue provides context for learning the target word's meaning.

4 Keyword Sequence Retrieval Validation

In this section, we validate each module of our keyword sequence retrieval system. We first describe our datasets for training and evaluation, then present detailed results for each module.

4.1 Dataset

Keyword Pool We construct a keyword candidate pool by filtering out non-lexical items such as grammatical particles, suffixes, and sentence-final endings from the Basic Korean Dictionary dataset (Ha, 2023).

The keyword pool consists of 55,316 unique entries that are phonologically representative and semantically well-formed.

Train We construct a training dataset of 2,870 English–Korean word pairs with aligned IPA transcriptions. English vocabulary items originate from standardized GRE preparation materials, including official Educational Testing Service guides and commercial resources (Magoosh, 2021; Princeton Review, 2020). Each entry in our dataset includes an English word, its Korean transliteration, and IPA transcriptions for both languages, with syllable-level boundaries annotated in the Korean IPA.

For English, we obtain IPA transcriptions directly from the Oxford University Press (n.d.), which provides standardized phonetic representations widely used in linguistic research.

For Korean, we first extract transliterations of English words from the Aha Dictionary (n.d.). These transliterations are segmented into syllable blocks of the Korean writing system (Hangul), each composed of an initial consonant, a medial vowel, and an optional final consonant. Hangul is often characterized as an alphabetic syllabary, where individual graphemes (jamo) form syllabic blocks (kulja) corresponding to single phonological units (Pastore, 2019). Each syllable block is then converted into its phonetic representation using the rule-based Hangul-to-IPA conversion method (Nam, 2022), which encodes standard Korean phonological processes (Shin et al., 2012). To identify syllable boundaries within the resulting IPA sequence, we extract the final IPA symbol from each block and annotate it with a binary indicator denoting syllable-final positions. This procedure enables consistent segmentation and alignment of syllable-level IPA representations across English and Korean.

Test We use the book KSS (Gyeong, 2020) as our baseline for human-authored verbal cues, designed for native Korean speakers learning English. The vocabulary targets advanced-level standardized tests, including government employee entrance exams, university transfer admissions, the TOEFL (Educational Testing Service), and the TEPS (Seoul National University). From this vocabulary, we construct a test set of 36 words.

4.2 Cross-lingual IPA Transliteration

We validate the module using two metrics: Character Error Rate (CER) and Exact Match Rate (EMR). CER quantifies the proportion of character-level errors, including insertions, deletions, and substitutions, between the predicted output and the reference. This metric effectively captures fine-grained phonological discrepancies, which is important for transliteration tasks involving languages with complex phonotactics or lacking clear word boundaries. EMR measures the percentage of outputs that exactly match the reference sequences. It serves as a strict criterion for evaluating whether the model produces completely accurate transliterations. Our model achieves a CER of 3.95% and an EMR of 75.56% for the train data set.

To understand how our model addresses phonological divergence between English and Korean, we analyze attention patterns (See Appendix Figure 6). The visualizations reveal the model's strategies for cross-linguistic challenges: English affricates decompose into multiple Korean consonants; compatible sounds maintain one-to-one mappings; English diphthongs expand to accommodate Korean's vowel inventory; and syllable structures adapt to Korean phonotactic constraints. These patterns confirm the model's ability to dynamically adjust its mapping strategy based on input characteristics.

4.3 Segmentation

We validate the module using boundary-level F1 score, which measures the model's precision and recall in identifying syllable boundaries. Since this module processes the output from the preceding transliteration component, we establish ground truth through manual annotation.

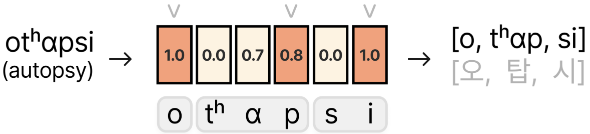


Figure 2: Visualization of the predicted syllable sequence of the English word *autopsy*. For /ot^{h}apsi/, the model assigns high boundary probabilities after o, p, and i, segmenting the sequence into [o, t^{h}ap, si].

Figure 2 illustrates the output of our model. For L1-adapted phoneme sequence /ot^{h}ɑspsi/ derived from the English word *autopsy*, the model assigns boundary probabilities that segment this sequence into phonologically valid Korean syllables.

Our model achieves a perfect boundary-level F1 score of 1.00 compared to reference boundaries on system transliteration outputs. This result indicates exceptional precision in identifying syllable junctures within the predicted L1-adapted sequence.

4.4 Keyword Matching

We validate the module by performing an ablation study on the structural alignment adjustment term to assess its contribution to our similarity function ϕ. The experimental results confirm that this component significantly improves syllable-level matching, which constitutes a critical factor for Korean phonological perception (Siew et al., 2021).

The comparison between retrieval methods shows clear differences in outcome quality. For the English word *demolish* (IPA: /dɪmalɪʃ/), the cosine similarity approach alone retrieves Korean keyword sequence with IPA transcriptions /pimilli/ and /ʃwi/, whereas our complete similarity function identifies Korean keyword sequence transcribed as /tɛmullim/ and /ʃwi/. Similarly, for *reckon* (IPA: /rɛkən/), the cosine-only method produces Korean keyword sequence with IPA representations /nɛ/ and /k^{h}an/, while our enhanced approach yields Korean keyword sequence represented as /lɛgɛ/ and /k^{h}ʌnsɛp/. These examples confirm that our weighting mechanism successfully prioritizes syllable matching as intended.

The structural alignment adjustment enables our model to identify keyword sequence that preserve syllabic structure and phonological patterns aligned with Korean perceptual tendencies, even when this preservation necessitates selection of candidates with slightly greater overall phonological distance.

5 Automated Evaluation

In the following sections, we detail how we evaluate the quality of the keyword sequence generated using PHONITALE, and we also discuss how we evaluate the quality of the generated verbal cues.

5.1 Dataset

We use same dataset mentioned in 4.1 for evaluation. We replicate Lee et al. (2024) in a same way for generating cross-lingual verbal cues, except that we translate prompts that were originally written in English with in-context examples for English-to-English learning into Korean, and use in-context examples from KSS to compare with PHONITALE. Hereafter, we refer to the human-authored book as **KSS**, Lee et al. (2024) as **OGR**, and PHONITALE as **PHT**. We use GPT-4o (temperature = 0.7) for both OGR and PHT throughout the entire pipeline to ensure fair comparison, following the temperature setting from Lee et al. (2024).

5.2 Metrics

5.2.1 Keyword Sequence

We evaluate quality on three aspects: phonetic similarity, keyword omission, and keyword modification. We evaluate phonetic similarity using our IPA-based contrastive model, which measures how closely the concatenated Korean keyword sequence resembles the phonetic form of the English word.

We define keyword omission as the proportion of proposed keyword sequence that are missing from the generated verbal cue relative to the total keyword count. Since the mnemonic method depends on combining multiple keywords to approximate the target word, omitting even one can disrupt the intended phonetic connection.

We also track keyword modification, which represents the ratio of keywords that appear in altered forms relative to the total keyword count. These modifications can shift pronunciation away from the target word and weaken the mnemonic link. (See Appendix Table 6 for examples.)

5.2.2 Verbal Cue

We evaluate the quality of verbal cue on *context completeness* and *perplexity* following Lee et al. (2024). Again, as we do not use the imageability metric for keyword sequence, we also do not calculate the imageability score of the verbal cue.

We calculate context completeness as in Section 3.2, while we calculate perplexity as a proxy for coherence, using KoGPT2-base-v2 (SKT, 2021), OpenAI's GPT-2 pretrained on large-scale Korean text data and adapted for natural language understanding and generation tasks in Korean.

5.3 Results

Table 1 shows that PHT achieves superior performance compared to other methods in the evaluation of both keyword and verbal cue quality.

5.3.1 Keyword Sequence

OGR, relying on LLMs for generating keyword sequence, frequently includes keywords that either do not exist in standard lexicons or lack everyday usage frequency. This results in substantial modifications when the keyword sequence is converted to verbal cues. Further, the modifications reduce the phonetic similarity with target English words.

KSS, authored by human, one possible reason for the low phonetic similarity is the its substitution of L1 meanings for L2 prefixes (e.g., *re-*, *in-*) and L2 suffixes (e.g., *-cracy*). For example, *re-* is mapped to the L1 word meaning *again*, and *in-* is mapped to the L1 word meaning *inside*.

5.3.2 Verbal Cue

OGR shows relatively lower performance in context completeness compared to other methods due to its excessive use of keywords. Since OGR focuses on splitting the target word into as many syllables as possible, the number of keywords corresponds to the number of syllables. Even with modified keywords not in standard lexicons or common use, it is generally difficult to generate natural and coherent context that effectively hints at the meaning of the target word.

For example, for the target word *frivolous*, OGR generates the keyword sequence /p^{h}urwn/ (blue), /pɑl/ (field), and /losw/ (Ross). The keyword sequence is shown in the verbal cue as "The reckless woman who was scolded by *Ross* in the *blue field*," with perplexity score 689.3. On the other hand, PHT retrieves the keyword sequence /p^{h}iri/ (flute) and /palladw/ (Ballad), shown as: "He, singing a *ballad* with the *flute*, acted rashly," with perplexity score of 231.0, which confirms that using two segments, as discussed in Section 3.1.2, achieves better performance in cross-lingual setting.

Following the highest perplexity score observed with OGR, which results from the use of unconventional keywords, KSS shows the next highest score. This is most likely due to the incorporation

Method	Phonetic$^{\uparrow}$	Omission$^{\downarrow}$	Modification$^{\downarrow}$	Context$^{\uparrow}$	Perplexity$^{\downarrow}$
KSS	0.74	3.7%	-	0.38	553.92
OGR	0.86	3.4%	24.8%	0.29	691.01
PHT	**0.95**	**0%**	**3.3%**	**0.39**	**433.41**

Table 1: Comparative analysis of metrics on keyword sequences and verbal cues. The keyword modifications of KSS was omitted because it does not provide information on the generation processes of keyword sequences and verbal cues.

Prompt	Context$^{\uparrow}$	Perplexity$^{\downarrow}$
OGR	0.29	490.13
PHT	**0.39**	**433.41**

Table 2: Comparison of verbal cue quality metrics using different prompt strategies while keeping all other components same as in the PHT system.

Model	Context$^{\uparrow}$	Perplexity$^{\downarrow}$
EXAONE3.5	0.38	450.30
GPT-4o (PHT)	**0.39**	**433.41**

Table 3: Ablation results on verbal cue generation using different models.

of L2 morphological elements in the keywords, as mentioned earlier. These incorporation introduce irregularities that make the model harder to predict, resulting in higher perplexity scores. Beyond their surface inclusion, KSS often requires learners to disambiguate polysemous morphemes such as *re-*, which can mean either *again* or *back* depending on the context. These inconsistencies in semantic interpretation and structural mapping increase irregularity in surface realizations, thereby hindering accurate verbal cue generation.

5.4 Ablation Studies

We conduct ablation studies on two key aspects: the prompts used for verbal cue generation and the models employed to generate these cues.

5.4.1 Prompts

Table 2 shows a result on ablating prompt for generating verbal cues. As discussed in Section 3.2, OGR uses two-step approach of generating a story then summarizing while PHT generates verbal cue right away. The result shows that PHT achieves better performance on both metrics, indicating that the two-step approach might be beneficial for English-only or verbal cue generations that require multiple keywords. However, in our cross-lingual setting, where there are only two keywords, generating verbal cue right away generates a better verbal cue.

5.4.2 Language Model

Table 3 shows a result on ablating language models for verbal cue generation. We utilize EXAONE3.5:32B, a 32-billion parameter open-sourced model with enhanced performance on Korean language tasks (Research et al., 2024) to test the language models suited for Korean language tasks can be an alternative for GPT4-o. The results shows that EXAONE3.5 achieves comparable performance on context completeness, and higher perplexity than GPT-4o. These results suggest that EXAONE-3.5 is a viable alternative, particularly when considering the cost and accessibility advantages of open-source models over proprietary ones.

6 Human Evaluation

6.1 Participants

We recruit Korean-native adults with intermediate English proficiency through university communities and LinkedIn. During the screening process, we also balance the participants' proficiency levels across groups. After screening, we assign a total of 51 individuals, with 17 in each of the experimental groups: KSS, OGR, and PHT (see Appendix Section D.1 for details).

6.2 Evaluation Setup

We design our evaluation to jointly assess short-term recall (Ellis and Beaton, 1993; Savva et al., 2014; Lee and Lan, 2023) and participant preference ratings (Lee et al., 2024), to measure whether the verbal cues are helpful and whether learners prefer them. We implement a web platform to conduct an experiment comprising learning, testing, and feedback phases, where the learning phase differs across groups by presenting keyword sequences and verbal cues specific to each condition.

6.3 Evaluation Procedure

Participants complete three rounds of learning and testing, consisting of 12 English words. In the learning phase, they are presented with the English word, its Korean meaning, audio pronunciation, Korean keyword sequence, and a verbal cue. The testing includes two tasks: recognition (recalling the meaning of the English word) and generation (producing the English word). In the feedback phase, participants rate each verbal cue on three aspects on a 5-point Likert scale: *helpfulness*, *coherence*, and *imageability* (See Appendix D.2 for procedure).

6.4 Metrics

6.4.1 Correctness

We assess the correctness of recognition and generation response using LLM-as-a-judge (GPT-4o) (Chiang and yi Lee, 2023). Previously, Savva et al. (2014) employ Levenshtein distance for assessing correctness. However, as the responses from recognition might involve synonym usage, minor part-of-speech variations, or unintentional typos, relying solely on surface-level string similarity metrics like Levenshtein distance may lead to misleading evaluations. Therefore, we adopt a more semantically aware approach by leveraging GPT-4o as a judge to assess the alignment between the model output and the answer (See Appendix D.4.1 for details).

6.4.2 Preference Ratings

We adopt the three criteria from Lee et al. (2024), except that we replaced usefulness with helpfulness to assess how much each cue aided memorization, rather than measuring usefulness in the absence of a recall test. Helpfulness measures how effective the cue is for memorizing the English word. Coherence measures the logical soundness of the verbal cue. Imageability measures how well the cue evokes vivid imagery in the participant's mind.

6.5 Results

6.5.1 Correctness

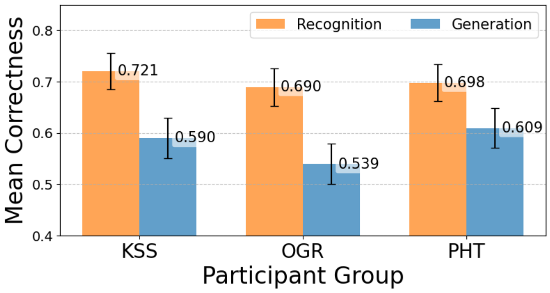


Figure 3: Mean correctness scores by participant group. Error bars indicate standard error.

Figure 3 shows correctness scores for both recognition and generation tasks across groups. In the recognition task, KSS achieves the highest correctness, followed by PHT and OGR. However, statistical tests indicate no significant differences between the groups. In the generation task, PHT achieves the highest correctness, followed by KSS and OGR. Analysis shows that PHT significantly outperforms OGR ($p < .05$), while the difference between PHT and KSS is not statistically significant.

These results highlight two key points. First, PHT performs comparably to human-authored cues, suggesting that LLM-generated prompts can be as effective as those created by humans. Second, our focus on phonetic alignment, rather than imageability, proves beneficial in the context of Korean-English vocabulary learning.

6.5.2 Preference Ratings

Group	Helpfulness	Coherence	Imageability
KSS	3.50 (1.41)	3.64 (1.42)	3.68 (1.39)
OGR	2.35 (1.40)	2.33 (1.37)	2.41 (1.41)
PHT	2.40 (1.26)	2.26 (1.21)	2.44 (1.23)

Table 4: Comparison of mean (standard deviation) of 5-point Likert scale participant ratings by group.

Table 4 shows preference ratings across groups. The KSS's ratings are statistically significantly higher than the others across all three criteria ($p < .001$), indicating that participants prefer human-authored cues over LLM-generated ones.

PHT receives higher ratings than OGR for helpfulness and imageability, while OGR is rated higher for coherence. However, these differences were not statistically significant. Notably, although PHT achieves significantly higher correctness during generation, this does not correlate with helpfulness. This finding is consistent with prior work showing that subjective preference does not always align with verbal cue effectiveness (Balepur et al., 2024). In terms of coherence, OGR receives higher ratings because its cue generation transforms a meaningless keyword into a meaningful word, as shown in high modification rate, providing greater flexibility and resulting in more logically coherent cues.

6.5.3 Case Study

PHT achieves higher correctness than KSS by generating keyword sequences that better preserve consonantal structure while maintaining phoneme-level alignment with the target word. For example, in words containing /r/ such as *reckon* and *render*, PHT selects initial keywords beginning with /l/, which is phonetically closer to /r/, whereas KSS selects /n/, resulting in less aligned mnemonics.

We assume that keyword sequences with stronger phonological alignment contribute more effectively to learners' ability to establish and retain accurate word associations.

However, KSS achieves higher correctness than PHT when keyword sequences are culturally rooted. For example, for *felon*, KSS adapts the idiom "to administer cudgel strokes" into the verbal cue *one who will cudgel-beat, therefore, a felon.* This construction is grammatically incorrect because it describes the one doing the beating rather than the one being beaten, yet learners readily reinterpret it as referring to the person who deserves punishment. The effectiveness of this vivid and culturally familiar cue is shown from its higher preference score compared to PHT. In contrast, PHT selects *Peleus*, a mythological name that preserves phonological alignment but lacks cultural resonance, making it harder to remember. This case illustrates how KSS benefits from culturally rooted and expressive forms, while current LLMs, constrained by grammaticality, struggle to produce such non-standard yet pedagogically effective cues.

7 Conclusion and Future Works

In this paper, we introduced PhoniTale, a novel system combining keyword sequence retrieval with verbal cue generation. Automated and human evaluations show that our approach performs comparably to human-authored cues and outperforms the method proposed by Lee et al. (2024). Furthermore, recall tests indicate our system achieves similar accuracy in recognition and statistically higher performance in generation. These results suggest that our strategy of leveraging phonetic similarity for mnemonic generation is effective.

Future extensions of this work are twofold. First, the system can be scaled to a broader range of typologically diverse languages, including syllable-timed languages such as Japanese and Spanish and tonal languages such as Mandarin and Vietnamese. For example, in English–Japanese, the word *render* pronounced as /rɛndər/ may be adapted as /renda/ in Japanese, segmented into [ren] and [da], and matched with native keywords such as レン "ren" (love) and だ "da" (copula) based on phonetic proximity. Provided that IPA-aligned L2–L1 transliteration data is available, the framework adapted to new language pairs. Additional refinements, such as language-specific syllable segmentation or tone modeling, can also enhance phonological compatibility across languages.

Second, we plan to extend the phoneme-anchored retrieval system to code-switched speech recognition. In such settings, phonological cues often transcend language boundaries, complicating identification of transliterated loanwords and domain-specific terms. By leveraging IPA-based representations, our approach offers a language-agnostic substrate for capturing cross-lingual phonetic similarity. This can improve recognition of borrowed or specialized vocabulary that deviates from canonical pronunciations, thereby reducing Word Error Rate (WER) in code-switched ASR scenarios. We plan to evaluate this by aligning phonetic units across typologically distant languages and assessing recognition gains for foreign-sounding or morphologically irregular tokens.

Together, these future directions aim to evolve PhoniTale into a more versatile, language-agnostic tool with applications spanning multilingual mnemonic generation, resource creation, and speech recognition in phonologically diverse or code-mixed environments.

Limitations

Our investigation exhibits four primary constraints. First, we limit our research scope to English-Korean language pairs due to the limited availability of training data, necessitating future adaptations for other language combinations with distinct phonological structures and orthographic systems. Second, our evaluation methodology assesses only short-term recall performance rather than longitudinal retention. Future research requires delayed post-tests to evaluate long-term memory consolidation and mnemonic durability. Third, our vocabulary selection derives from standardized test materials targeting advanced-level English learners, potentially limiting PhoniTale's applicability for beginning learners acquiring common vocabulary. Fourth, while our Korean dictionary dataset includes some conjugated forms, its lexical coverage remains limited. The absence of commonly used loanwords, neologisms, and other everyday variants reduces the pool of potential keywords and constrains the naturalness of generated verbal cues.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Global Data-X Leader HRD program grant funded by the Korea government (MSIT) (IITP-RS-2024-00440626)

Ethics Statement

We conduct all human participant experiments in accordance with ethical guidelines after obtaining approval from our Institutional Review Board (IRB). Prior to participation, we present subjects with comprehensive consent forms detailing potential risks, anticipated benefits, expected time commitment, experimental activities, and compensation structure. All participants provide informed consent by acknowledging these terms before experiment commencement. The compensation structure aligns with local wage standards, with participants receiving between KRW 9,000 and 12,000 (approximately USD 6.50-8.70) based on the duration and complexity of their participation.

References

Aha Dictionary. n.d. Aha dictionary. http://aha-dic.com/. Accessed: 2025-04-15.

Richard C. Atkinson and Michael R. Raugh. 1975. An application of the mnemonic keyword method to the acquisition of a russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory*, 1(2):126–133.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Nishant Balepur, Matthew Shu, Alexander Hoyle, Alison Robey, Shi Feng, Seraphina Goldfarb-Tarrant, and Jordan Boyd-Graber. 2024. A SMART Mnemonic Sounds like "Glue Tonic": Mixing LLMs with Student Feedback to Make Mnemonic Learning Stick. *Preprint*, arXiv:2406.15352.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *Preprint*, arXiv:2305.01937.

CMU. 1993. Free pronouncing dictionary of english. http://www.speech.cs.cmu.edu/cgi-bin/cmudict?stress=-s&in=CITE. Accessed: 2025-05-17.

Robert Daland and Kie Zuraw. 2013. Does Korean defeat phonotactic word segmentation? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 873–877, Sofia, Bulgaria. Association for Computational Linguistics.

Educational Testing Service. Toefl® test - test of english as a foreign language. https://www.ets.org/toefl. Accessed: 2025-05-06.

Nick C Ellis and Alan Beaton. 1993. Psycholinguistic determinants of foreign language vocabulary learning. *Language learning*, 43(4):559–617.

Sunsik Gyeong. 2020. *KSS English Vocabulary: Government employee entrance exams, University transfer admissions, the TOEFL, and the TEPS Revised Edition*. KSSEDU, Seoul. Revised Edition.

Chris Ha. 2023. Basic korean dictionary. https://huggingface.co/datasets/hac541309/basic_korean_dict. Accessed: 2025-05-16.

International Phonetic Association. 1949. *The Principles of the International Phonetic Association: Being a Description of the International Phonetic Alphabet and the Manner of Using It, Illustrated by Texts in 51 Languages*. International Phonetic Association, London. Supplement to Le Maître Phonétique, January-June 1949.

Yoonjung Kang. 2003. Perceptual similarity in loanword adaptation: English postvocalic word-final stops in korean. *Phonology*, 20(2):219–273.

Yoonjung Kang. 2014. Voice onset time merger and development of tonal contrast in seoul korean stops: A corpus study. *Journal of Phonetics*, 45:76–90.

Yoonjung Kang, Jessamyn Schertz, and Sungwoo Han. 2022. *The Phonology and Phonetics of Korean Stop Laryngeal Contrasts*, page 215–247. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Michael Kenstowicz. 2005. The phonetics and phonology of korean loanword adaptation.

Kyumin Kim and Alexei Kochetov. 2011. Phonology and phonetics of epenthetic vowels in loanwords: Experimental evidence from korean. *Lingua*, 121:511–532.

Chang H Lee and Marcus Taft. 2017. Syllable-based phonological processes in korean word recognition. *Language and Cognitive Processes*, 32(1):1–20.

Jaewook Lee and Andrew Lan. 2023. Smartphone: Exploring keyword mnemonic with auto-generated verbal and visual cues. In *International Conference on Artificial Intelligence in Education*, pages 16–27. Springer.

Jaewook Lee, Hunter McNichols, and Andrew Lan. 2024. Exploring automated keyword mnemonics generation with large language models via overgenerate-and-rank. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5521–5542, Miami, FL, USA.

Jaewook Lee, Alexander Scarlatos, and Andrew Lan. 2025. Interpretable mnemonic generation for kanji learning via expectation-maximization. *arXiv preprint arXiv:2507.05137*.

Minchul Lee. 2020. Corpus for sentiment analysis. https://github.com/bab2min/corpus/tree/master/sentiment. Accessed: 2025-05-16.

Magoosh. 2021. *GRE Prep by Magoosh*. Magoosh Online Test Preparation.

Connor Mayer and Max Nelson. 2020. Syllable detection and segmentation using temporal convolutions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3704–3713.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

mphilli. 2018. eng-to-ipa: English to ipa transcription library. https://pypi.org/project/eng-to-ipa/. Python package that utilizes the Carnegie-Mellon University Pronouncing Dictionary.

Stanley Nam. 2022. hangul_to_ipa: Korean hangul to ipa converter. https://github.com/stannam/hangul_to_ipa. Accessed: 2025-04-17.

OpenAI. 2024. Hello gpt-4o. Accessed: 2025-02-19.

Oxford University Press. n.d. Oxford english dictionary. https://www.oed.com/?tl=true. Accessed: 2025-03-10.

Changho Park and Ping Li. 2009. *Visual processing of Hangul, the Korean script*, page 379–389. Cambridge University Press.

Martina Pastore. 2019. Processing an alphabetic syllabary: Investigating the orthographic code for korean. Master's thesis, KU Leuven, ESAT, PSI-Speech, Master of Science in Artificial Intelligence (Speech and Language Technology). Master's thesis; thesis supervisor: Prof. dr. Dirk Van Compernolle.

Princeton Review. 2020. *Cracking the GRE*. Princeton Review.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Sihoon Yang, Heuiyeen Yeen, and Hyeongu Yun. 2024. Exaone 3.5: Series of large language models for real-world use cases. *Preprint*, arXiv:2412.04862.

Manolis Savva, Angel X. Chang, Christopher D. Manning, and Pat Hanrahan. 2014. Transphoner: Automated mnemonic keyword generation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, pages 3725–3734.

Seoul National University. Teps - test of english proficiency developed by seoul national university. https://www.teps.or.kr. Accessed: 2025-05-06.

Jiyoung Shin, Jieun Kiaer, and Jaeeun Cha. 2012. Phonological rules of korean (i). In *The Sounds of Korean*, chapter 8. Cambridge University Press, Cambridge, UK. Published online: 05 November 2012.

Cynthia S.Q. Siew, Kwangoh Yi, and Chang H. Lee. 2021. Syllable and letter similarity effects in korean: Insights from the korean lexicon project. *Journal of Memory and Language*, 116:104170.

SKT. 2021. Kogpt2: Korean gpt-2 pretrained cased. https://github.com/SKT-AI/KoGPT2. Accessed: 2025-05-16.

Jin-Young Tak. 2012. Variations in english loanword adaptation in korean phonology: Attributable to the borrowing language's internal grammar or speech perception? *The Journal of English Language and Literature*, 58:541–563.

Hyesook Yoon and Donald J Bolger. 2015. Perceptual similarity in korean word recognition: An analysis of onset and rime similarities. *Journal of Cognitive Psychology*, 27(6):736–750.

Appendix

A Introduction

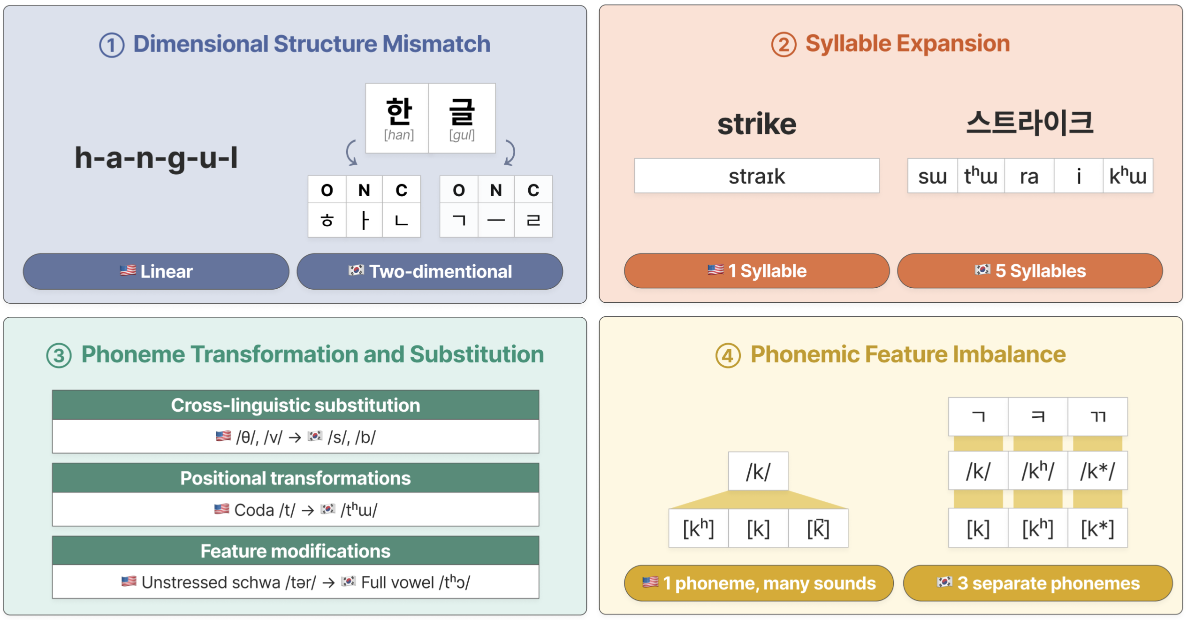


Figure 4: Four key challenges in English-Korean phonological alignment: (1) Dimensional structure mismatch: Korean's two-dimensional syllabic blocks versus English's linear sequence. (2) Syllable expansion due to consonant cluster resolution. (3) Phoneme transformation: Korean lacks certain English distinctions while English lacks Korean's three-way consonant contrast. (4) Phonemic contrast differences: Korean's systematic three-way distinction versus English's position-dependent allophones.

B Methodology

B.1 PHONITALE Architecture

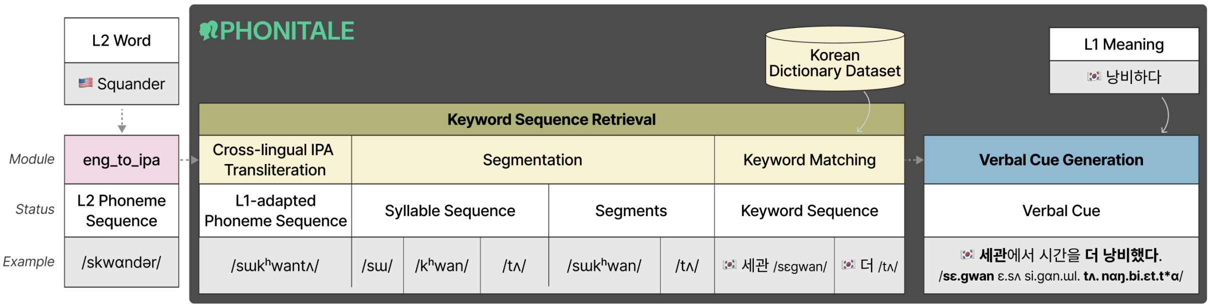
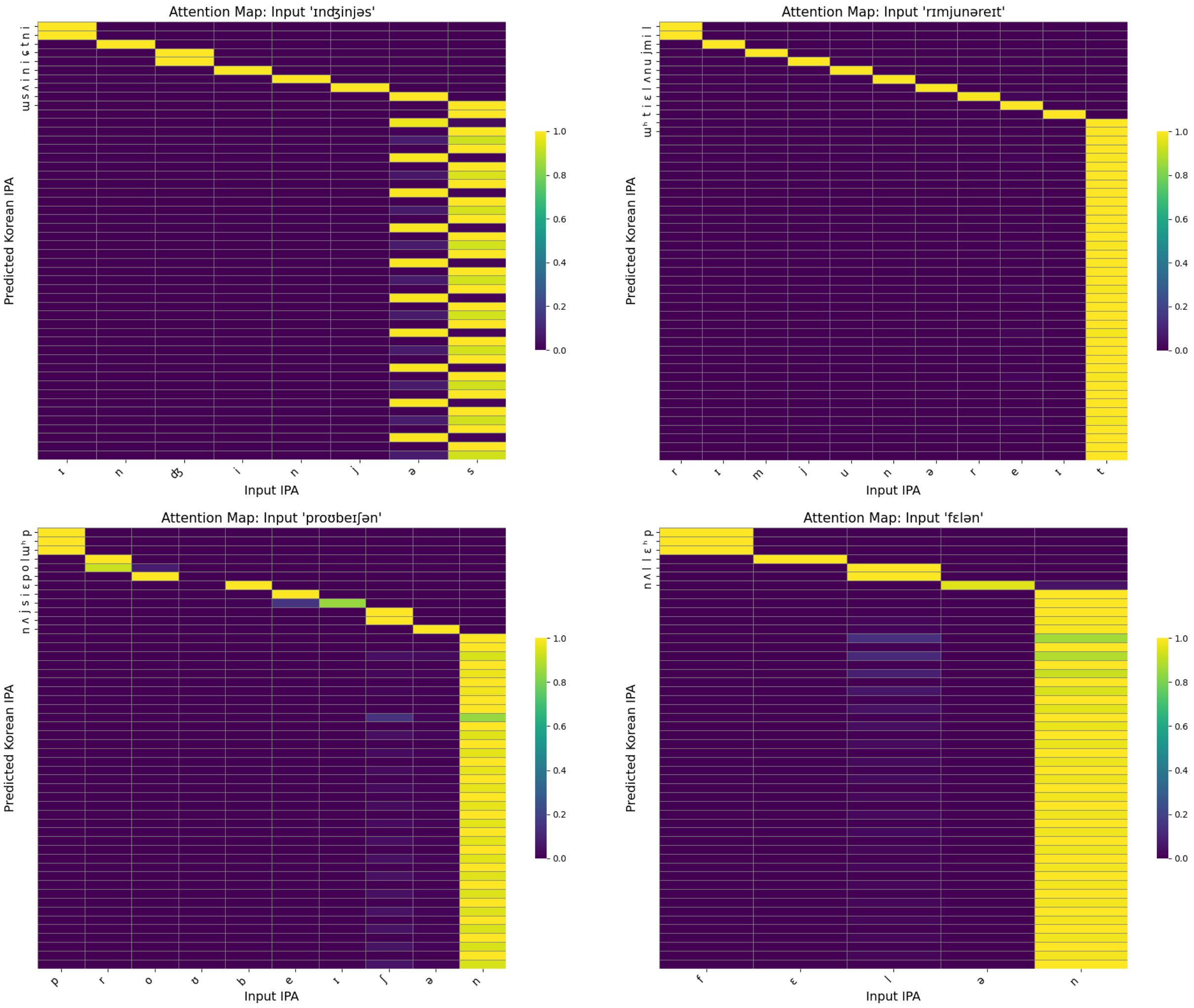


Figure 5: We demonstrate this two-phase pipeline through a running example of w_{L2} "squander". The system first converts the word into P_{L2} (/skˈwandər/) using the eng-to-ipa library (mphilli, 2018), which is based on the CMU Pronouncing Dictionary (CMU, 1993). The system then generates $\widehat{P}_{\mathrm{L1}}$ (/sɯk$^{\mathrm{h}}$wantʌ/), predicts syllable sequence (/sɯ/, /k$^{\mathrm{h}}$wan/, /tʌ/), and derives the segments (/sɯk$^{\mathrm{h}}$wan/, /tʌ/). The system retrieves $\mathcal{W}_{\mathrm{L1}}$ with IPA transcriptions /sæ.gwan/ and /tʌ/, and uses them to construct a verbal cue: "**sɛ.gwan** ɛ.sʌ si.gan.ɯl. **tʌ naŋ.bi.ɛt.t*a**" (English translation: **Wasted more** time at **customs**).

B.2 Syllable Prediction


Attention Map: Input 'ɪnʤinjəs'

Attention Map: Input 'rɪmjunəreɪt'

Attention Map: Input 'proʊbeɪʃən'

Attention Map: Input 'fɛlən'

Figure 6: Representative attention maps visualizing cross-lingual phonological alignment patterns. **Top left**: Attention map for /ɪndɪdʒəniəs/ (indigenous) showing affricate decomposition where English affricates are mapped to multiple Korean consonants, creating vertical attention patterns. **Top right**: Attention map for /rɪmjunəreɪt/ (remunerate) demonstrating near one-to-one alignment with strong diagonal attention patterns where phonological structures are compatible across languages. **Bottom left**: Attention map for /proʊbeɪʃən/ (probation) highlighting diphthong expansion and coda realignment where English diphthongs spread across multiple Korean vowels and final consonants shift to match Korean syllable constraints. **Bottom right**: Attention map for /fɛlən/ (felon) illustrating structure-induced elongation where the short CVC English syllables must expand to fit Korean's more restrictive syllable templates, creating distributed attention across additional segments.

B.3 Keyword Matching

```
Require: Segment $S_i$, Candidate keyword $w^*_{\mathrm{L1},i}$
Require: Embedding function $\mathbf{v}(\cdot)$
Require: Parameters: $\lambda_{\text{syll}}, \lambda_{\text{first}}, \lambda_{\text{substr}}, \lambda_{\text{early}}$
score ← $\cos\Big(\mathbf{v}(S_i), \mathbf{v}(w^*_{\mathrm{L1},i})\Big)$        ▷ Base similarity
if SyllableOverlap($S_i$, $w^*_{\mathrm{L1},i}$) then
    if InitialSyllableMatch($S_i$, $w^*_{\mathrm{L1},i}$) then
        $score \leftarrow score + \lambda_{\text{syll}} \times \lambda_{\text{first}}$
    else
        $score \leftarrow score + \lambda_{\text{syll}}$
    end if
else
    if SubstringInclusion($S_i$, $w^*_{\mathrm{L1},i}$) then
        $score \leftarrow score + \lambda_{\text{substr}}$
    else if EarlyPhoneMatch($S_i$, $w^*_{\mathrm{L1},i}$) then
        $score \leftarrow score + \lambda_{\text{early}}$
    end if
end if
return $score$
```

Algorithm 1: Keyword scoring algorithm used in the keyword matching module. We apply cosine similarity as the base score, and augment it with structural alignment adjustments, which are empirically tuned on the development set: $\lambda_{\text{syll}} = 0.9$ (syllable overlap bonus), $\lambda_{\text{first}} = 2.0$ (initial-syllable match multiplier), $\lambda_{\text{substr}} = 0.3$ (substring inclusion bonus), and $\lambda_{\text{early}} = 0.2$ (early-phone match bonus). These additions are designed to enhance alignment with syllable-based perception patterns in Korean, aiding memorability and cue effectiveness.

B.4 PHONITALE Prompt

The prompt was originally designed in Korean. For reproducibility, we provide both the original and its English translation.

Prompt	게임 이름: 이야기 엮기 놀이 *Game name: Story-Chaining Game* 게임 설명: 이야기 엮기 놀이에서 플레이어들은 목표 단어 후보와 키워드 세트를 받습니다. 플레이어들의 임무는 이 단어들을 교묘하게 사용하여 짧고 간결한 한 문장의 이야기를 만드는 것입니다. 궁극적인 도전은 목표 단어 후보 중 하나를 선택적으로 포함하고 제시된 순서대로 정확히 키워드를 포함하는 한 문장의 이야기를 구성하는 것입니다. *Game description: In the Story-Chaining Game, players receive a target word candidate set and a keyword set. Their task is to craft a short, concise, single-sentence story that cleverly incorporates these words. The ultimate challenge is to construct a sentence that includes at least one of the target word candidates of your choice and strictly uses the provided keywords in the given order.* 게임 규칙/제약사항: *Game rules/constraints:* 1. 각 플레이어는 목표 단어 후보와 키워드 세트를 받습니다. *1. Each player receives a set of target word candidates and keywords.*

2. 목표 단어를 먼저 결정해야 합니다. 목표 단어 후보가 하나라면, 그 단어가 곧 목표 단어가 됩니다. 목표 단어 후보가 여러 개라면, 플레이어는 그 중 하나를 선택해야 합니다.

2. The target word must be chosen first. If there is only one candidate, that word becomes the target word. If multiple candidates are provided, the player must select one.

3. 목표 단어와 키워드를 사용하여 한 문장으로 된 짧은 이야기를 만들어야 합니다.

3. The target word and keywords must be used to create a short, single-sentence story.

4. 키워드는 주어진 순서대로 정확히 등장해야 합니다.

4. Keywords must appear exactly in the specified order.

5. 한 문장의 이야기에는 목표 단어가 포함되어야 하며, 한 번만 나타나야 합니다. 목표 단어는 꺾쇠 괄호(< >)로 묶어 강조해야 합니다.

5. The story must include the target word exactly once, and it should be highlighted using angle brackets (< >).

6. 전체 내용은 json 형식으로 반환해야 합니다.

6. The entire output must be returned in JSON format.

7. 플레이어는 키워드의 순서를 재배열하는 것이 엄격히 금지됩니다.

7. Rearranging the order of the keywords is strictly prohibited.

다음은 입력과 출력이 어떻게 보여야 하는지에 대한 예시입니다:

The following are examples of the expected input and output format:

[Input]

목표 단어 후보: <취소하다>

Target word candidates: <countermand>

키워드 세트: 카운터, 만두

Keyword set: /k^{h}a.un.t^{h}ʌ/ (counter), /man.du/ (dumpling)

[Output]

{

"목표 단어": "취소하다",

"target word": "countermand",

"이야기": "그는 카운터에서 만두 주문만 <취소했다>."

"story": "He <countermanded> the dumpling order at the counter."

}

[Input]

목표 단어 후보: <범인, 범죄자>

Target word candidates: <culprit, criminal>

키워드 세트: 칼, 뿌리다

Keyword set: /k^{h}al/ (knife), /[p^{h}u.ri.da]/ (scatter)

[Output]

{

	"목표 단어": "범죄자",
	"target word": "culprit",
	"이야기": "칼을 뿌리는 <범죄자>."
	"story": "The <culprit> scattered knives."
	}
	[Input]
	목표 단어 후보: <튀기다, 첨벙거리다>
	Target word candidates: <fry, splash>
	키워드 세트: 수풀, 쉬
	Keyword set: /su.p^{h}ul/ (bush), /çi/ (pee)
	[Output]
	{
	"목표 단어": "튀기다",
	"target word": "splash",
	"이야기": "수풀에 쉬를 하다 물을 <튀겼다>."
	"story": "While peeing in the bush, water <splashed>."
	}
	[Input]
	목표 단어 후보: <나태한, 게으른>
	Target word candidates: <sluggish, lazy>
	키워드 세트: 인어, 덜렁대다
	Keyword set: /in.ʌ/(mermaid), /tʌl.lʌŋ.dɛ.da/ (fumble)
	[Output]

Response	{
	"목표 단어": "게으른",
	"target word": "indolent",
	"이야기": "인어는 덜렁대며 <게으르게> 움직였다."
	"story": "The mermaid fumbled around and moved indolently."
	}

Table 5: Prompts for generating verbal cues.

C Automatic Evaluation

Issue Type	**Target Word**	**Proposed Keyword Sequence**	**Used Keyword Sequence**	**Description**
Omission	provisional	p^{h}wro, pisʌ, nʌl	p^{h}wro, pisʌ	The keyword nʌl is omitted from the verbal cue.
Modification	reticent	lɛ, t^{h}i, sɛnt^{h}w	lɛswt^{h}oraŋ, t^{h}i, sɛnt^{h}w	The keyword lɛ is modified to ləswt^{h}oraŋ.

Table 6: Examples of Keyword Omission and Modification

D Human Evaluation

D.1 Participants

We summarize the participant recruitment, screening, and group assignment process in Table 7. Vocabulary familiarity was assessed using a 12-word survey. Words were grouped into difficulty tiers and scored (3=High, 2=Medium, 1=Low) based on the percentage of participants who reported familiarity (see Table 9). Final group assignment ensured balance in vocabulary familiarity, age, and education level.

Step	**Description**
Recruitment	167 Korean-native adults via university communities and LinkedIn
Screening Task	12-word self-report survey (SAT/TOEFL/GRE vocabulary)
Scoring Method	Score = sum of recognized words weighted by difficulty (3 = high, 2 = medium, 1 = low)
Filtering	Top outliers removed using upper quartile threshold; bottom excluded if score ≤ 2 → 132 eligible participants
Group Assignment	Random assignment to KSS, OGR, and PHT with matched familiarity levels
Experiment Completion	55 participants completed the main task
Quality Filtering	Bottom 4 participants excluded based on lowest completeness scores
Final Sample	51 participants (17 per group)
Equivalence Check	No significant differences across groups: • Vocabulary familiarity ($p = .4378$) • Age ($p = .9100$) • Education level ($p = .3599$)

Table 7: Summary of participant recruitment, screening, and group assignment process.

Group	**Mean Age**	**Std Dev**	**Min Age**	**Max Age**
KSS	28.24	6.47	20	41
OGR	28.41	8.54	18	54
PHT	28.47	5.56	18	35

Table 8: Age statistics by group

Target Word	Difficulty Level	Assigned Score	Familiarity
intransigent	High	**3**	1.8%
reinstate	High	**3**	6.0%
horrendous	High	**3**	13.8%
sanction	High	**3**	21.0%
abdominal	Medium	**2**	25.1%
uphold	Medium	**2**	38.9%
deduce	Medium	**2**	42.5%
mutable	Medium	**2**	46.7%
hygiene	Low	**1**	54.5%
criterion	Low	**1**	55.7%
inverse	Low	**1**	78.4%
align	Low	**1**	81.4%

Table 9: Twelve English words used in the screening survey, grouped by assigned difficulty level and annotated with the percentage of participants who reported familiarity. These values serve as the basis for scoring vocabulary familiarity across participants.

D.2 Evaluation Procedure

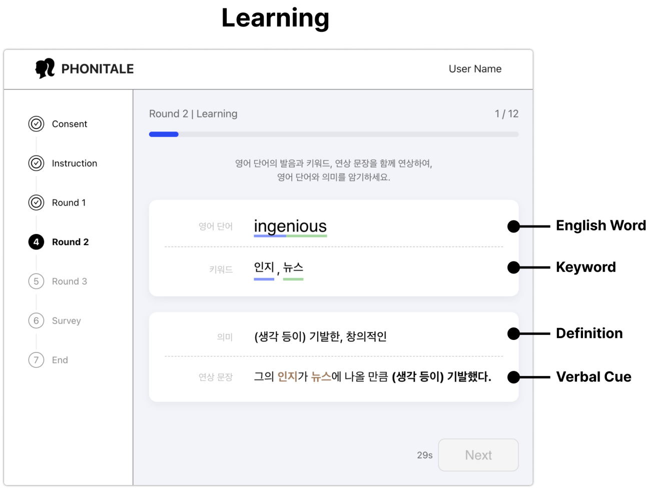
D.2.1 Procedure

Phase	Description
Learning	• For each word, participants see: – English word (visually segmented) – Korean definition – Audio pronunciation (played at 2s and 7s) – Korean keyword sequence – Verbal cue • Color underlining highlights phonological alignment • 30-second time limit (advance allowed after 15s) • 1-second blank screen between words
Testing – Recognition	• Task: Type the Korean definition • For each word: – English word – Audio pronunciation (played at 2s and 7s) • 30-second time limit • 1-second blank screen between words • Responses are used to compute correctness scores

Testing – Generation	• Task: Type the English word • For each word: – Korean definition • 30-second time limit • 1-second blank screen between words • Responses are used to compute correctness scores
Feedback	• Participants rate each mnemonic on three 5-point Likert scales: – Helpfulness: Cue supports recall of the word's meaning – Coherence: Sentence is logical and grammatically natural – Imageability: Cue evokes a vivid and concrete image

Table 10: Detailed task structure for learning, testing, and feedback phases. Each round includes 12 English words, repeated across three rounds (36 total).

D.2.2 Web User Interface



Round 2 | Learning 1 / 12

영어 단어의 발음과 키워드, 연상 문장을 함께 연상하여,
영어 단어와 의미를 암기하세요.

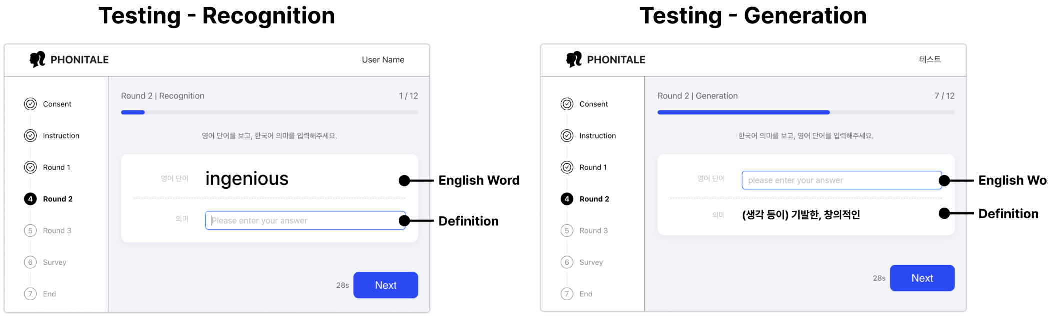
영어 단어 ingenious

키워드 인지 , 뉴스

의미 (생각 등이) 기발한, 창의적인

연상 문장 그의 인지가 뉴스에 나올 만큼 (생각 등이) 기발했다.

Figure 7: User interface for the learning phase. Each screen presents the English word, Korean definition, phonologically aligned keyword sequence, and a verbal cue.



Round 2 | Recognition 1 / 12

영어 단어를 보고, 한국어 의미를 입력해주세요.

영어 단어 ingenious

의미 Please enter your answer

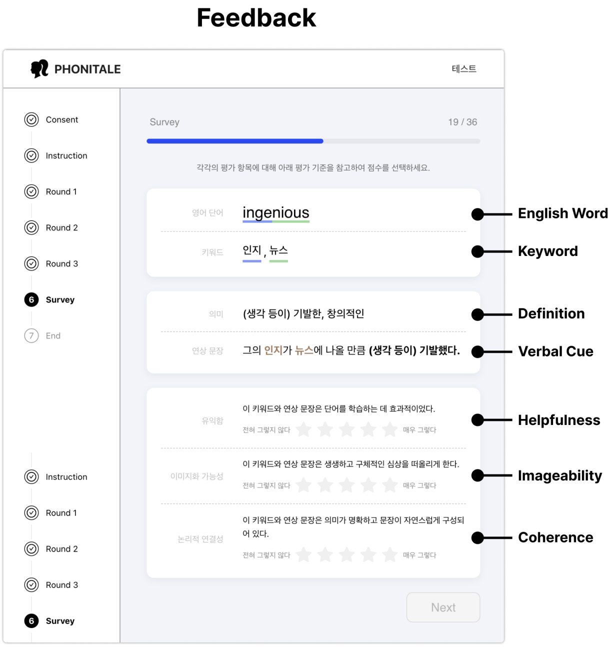
Round 2 | Generation 7 / 12

한국어 의미를 보고, 영어 단어를 입력해주세요.

영어 단어 please enter your answer

의미 (생각 등이) 기발한, 창의적인

Figure 8: User interface for the testing phase. Left: recognition task, Right: generation task.



Survey 19 / 36

각각의 평가 항목에 대해 아래 평가 기준을 참고하여 점수를 선택하세요.

영어 단어: ingenious

키워드: 인지 , 뉴스

의미: (생각 등이) 기발한, 창의적인

연상 문장: 그의 인지가 뉴스에 나올 만큼 (생각 등이) 기발했다.

유익함: 이 키워드와 연상 문장은 단어를 학습하는 데 효과적이었다.
전혀 그렇지 않다 ☆ ☆ ☆ ☆ ☆ 매우 그렇다

이미지화 가능성: 이 키워드와 연상 문장은 생생하고 구체적인 심상을 떠올리게 한다.
전혀 그렇지 않다 ☆ ☆ ☆ ☆ ☆ 매우 그렇다

논리적 연결성: 이 키워드와 연상 문장은 의미가 명확하고 문장이 자연스럽게 구성되어 있다.
전혀 그렇지 않다 ☆ ☆ ☆ ☆ ☆ 매우 그렇다

Figure 9: User interface for the feedback phase.
Participants rate each cue on helpfulness, imageability, and coherence using a 5-point scale.

D.2.3 Criteria

Guidelines for rating the 5-point Likert scale used to evaluate *helpfulness*, *coherence*, and *imageability* of verbal cues.

Scale	Explanation
High (5)	구조적으로 잘 연관되어 있고, 반복 학습 없이도 단어의 의미를 쉽게 떠올릴 수 있음 *Well-structured and strongly linked; supports effortless recall without repeated study*
Medium (3)	단서와 의미 사이에 약한 연결 고리가 있으나, 기억에 오래 남기엔 부족함 *Somewhat related, but not strongly memorable or sufficient for long-term recall*
Low (1)	단어의 뜻과 단서 사이에 직접적 연결이 거의 없어 기억하거나 학습하는 데 실질적인 도움이 되지 않음 *Weak or no clear link between cue and word meaning; offers little learning support*

Table 11: Instructions for rating **helpfulness** of the verbal cues.

Scale	Explanation
High (5)	논리, 어휘, 의미 흐름이 매끄럽고 자연스럽게 구성됨 *Logically coherent and lexically natural; overall meaning flows well*
Medium (3)	비교적 자연스럽지만, 문법이나 논리 흐름에서 약간 부자연스러움 *Fairly natural but with some grammatical or logical awkwardness*
Low (1)	문장이 어색하고 단어 해석과 연결성이 부족함 *Grammatically awkward or incoherent; lacks clarity or connection to the word*

Table 12: Instructions for rating **coherence** of the verbal cues.

Scale	Explanation
High (5)	익숙한 이미지로 쉽게 시각화되며 장면이 구체적으로 떠오름 *Evokes vivid and specific scenes using familiar imagery*
Medium (3)	단어와 관련된 이미지가 조금 있으나 모호하거나 약함 *Partially imageable, but vague or weakly related to the word*
Low (1)	장면이나 상황이 전혀 그려지지 않음 *Difficult or impossible to visualize any meaningful scene or context*

Table 13: Instructions for rating **imageability** of the verbal cues.

D.3 Evaluation Materials by Group

D.3.1 English Words Set

English Word	Korean Definition
albeit	비록 ∼이기는 하나
annihilate	전멸시키다, 붕괴하다
canny	약삭빠른, 영리한
esoteric	비법을 이어받은, 소수만 이해하는
incumbent	의무로서 지워지는, 현직의, 재임 중인
insolvent	파산한
meddlesome	참견하기를 좋아하는
probation	집행유예, 보호관찰, 수습
reckon	(∼라고) 생각하다, 예상하다
refurbish	(방, 건물 등을) 새로 꾸미다, 쇄신하다
resuscitate	소생시키다
upheaval	대변동, 격변
anachronism	시대착오, 시대착오적인 사람(관습생각)
bureaucracy	관료정치, 관료체제, 관료
delirium	정신 착란, 헛소리

demolish	때려 부수다, 파괴하다, 철거하다
felon	중죄인, 흉악범
incarcerate	투옥하다, 감금하다
ingenious	(생각 등이) 기발한, 창의적인
recidivist	상습범
redoubtable	가공할, 무시무시한, 경외할 만한
remunerate	보상하다, 보수를 주다
render	(어떤 상태가 되게) 만들다, ~하게 하다, 주다, 제출하다
repercussion	영향, 파급효과
autopsy	(사체의) 부검
congenital	타고난, 선천적인
fictitious	허구의, 지어낸
inebriate	취하게 하다, 술꾼
insurrection	폭동, 반란
intransigent	비협조적인, 비타협적인
inveterate	(습관 등이) 뿌리 깊은, 고질적인
mayhem	대혼란, 아수라장
peccable	과오를 범하기 쉬운
provisional	임시의, 일시적인
reimburse	갚다, 상환하다
squander	낭비하다

Table 14: English words used in the experiment with their corresponding Korean definitions.

D.3.2 KSS Keyword and Verbal Cue

The 36 English words used for human-authored mnemonics were selected from 경선식 영단어 공편토 (Gyeong, 2020). These selections are used solely for research and evaluation purposes.

English Word	KSS Keyword	KSS Verbal Cue
albeit	모두, B (all, B)	비록 학점이 모두 B이기는 하나 다음에는 A를 받을 것이다. (Although all grades are B, one will get an A next time.)
annihilate	언, 아이, 끄집어내다 (frozen, child, rescue)	얼음 속에 언 아이를 밖으로 끄집어내려고 얼음을 붕괴하다 (To rescue a frozen child outside, the ice must be annihilated.)
canny	캐니 (dig up)	선거에서 상대방 약점을 캐니? 즉, 약삭빠른, 영리한 (Do you "dig up" the opponent's weakness? That is, shrewd and clever.)

esoteric	애써, 때리 (to strive, to hit)	요리의 대가가 제자에게 "애 써!" 하고 때리며 자식에게만 가르쳐주는 비법 (A master chef says "Work hard!" while hitting the apprentice, teaching secret skills only to his own child.)
incumbent	수입, 번 (income, earned)	수입을 버는 것은 가장의 의무이기에 짤리지 말고 현직에 있어야 한다 (Earning income is a duty of the head of household, so one must remain in office.)
insolvent	안, 쌀, 타다 (in, rice, burnt)	곳간 안에 쌀이 다 타버려 파산한 (In the granary, all the rice has burnt, leaving one insolvent.)
meddlesome	중간, 몇 (middle, some)	사람들 사이 중간에 몇 번씩 나타나 참견하기 좋아하는 (Someone who repeatedly appears in the middle of people, eager to meddle.)
probation	풀어, 뵈이션 (to release, to see)	풀어주고 또 범행을 하는지 뵈이도록 지켜보는 것. 즉 집행유예, 보호관찰 (Release someone and see whether they commit another crime. That is, probation, parole.)
reckon	내 껀 (mine)	이 금도끼는 내 껀 아니라고 생각하다 (Thinking "this golden axe is not mine.")
refurbish	다시, 퍼, 빛이 (again, to scoop, light is)	다시 물을 퍼서 빛이 나게 씻어 새로 꾸미다 (Scoop water again to wash and make it shine — to refurbish.)
resuscitate	다시, 서시다 (again, stand up)	죽은 사람이 다시 일어서시게 하다. 즉 소생시키다 (To make a dead person rise again — to resuscitate.)
upheaval	위, 엉덩이 (up, hip)	거인이 엉덩이를 위로 들고 방귀를 세게 뀌자 사방이 진동하며 일어나는 대변동 (When a giant lifts up his hip and farts loudly, causing upheaval with vibrations everywhere.)
anachronism	아내, 끌어(내다)니즘 (wife, to drag out-ism)	칠거지악을 범했다며 아내를 집에서 끌어내는 시대착오적인 생각 (The anachronistic idea of dragging one's wife out of the house, claiming she committed the seven grounds for divorce.)
bureaucracy	행정부, 크라시 (govern, meaningless word) (bureau, govern)	의회나 정당이 아닌 행정부로(행정부의 관료들이) 통치하는 관료정치 (Bureaucracy where the executive branch (bureaucrats) govern instead of the parliament or political parties.)
delirium	닐리리아 (nil-li-ri-a)	"닐리리아" 하며 미쳐서 노래 부르는 정신착란 (Mental delirium of singing "nil-li-ri-a" while going crazy.)

demolish	뒤, 말리시다 (behind, stop)	술에 취해 물건들을 때려 부수는 사람을 뒤에서 말리시다 (Stop someone from behind who is drunkenly demolishing things.)
felon	팰, 놈 (to beat, guy)	곤장을 팰 놈, 즉 중죄인, 흉악범 (A guy to be beaten with a rod, that is, a felon, a criminal.)
incarcerate	안, 칼, 쓰래이! (in, cangue, put on!)	"감옥 안에서 칼을 써!" 하고 투옥하다, 감금하다 ("Put on the cangue in prison!" while incarcerating, confining.)
ingenious	안, 지니다 (in, to possess)	머리 안에 지녔어, 기발한 창의적인 생각을 (Possessing ingenious, creative thoughts in one's head.)
recidivist	다시, 씨디, 비슷 (again, CD, similar)	불법복제로 처벌받은 후 re(다시) CD를 비슷하게 불법복제하는 상습범 (A recidivist who, after being punished for illegal copying, again makes similar illegal CD copies.)
redoubtable	다시, 다, 울어 (again, all, cry)	영화에서 무시무시한 귀신이 나와 아이들이 다시 다 울어 (In the movie, a redoubtable ghost appears and makes all the children cry again.)
remunerate	뒤, 물어내다 (again, to pay back)	어떤 대가로 되돌려 물어내다. 즉 보상하다 (To pay back in return for something — to remunerate.)
render	낸다 (to pay)	심부름센터에 돈을 낸다. 그리고 하게 하다 (Pay money to an errand center. And to render/make something.)
repercussion	뒤, 퍼지다, 쿠션 (back, spread, cushion)	공이 뒤로 튀며 쿠션 효과를 내는 파급효과 (The repercussion effect of a ball bouncing back with a cushion effect.)
autopsy	오!, 톱, 보다 (oh!, saw, see)	오! 톱으로 시체를 잘라 자세히 보는 부검 (Oh! An autopsy where a corpse is cut with a saw to see in detail.)
congenital	큰, 제니, 털 (big, Jenny, hair)	제니라는 사람의 얼굴에 난 큰 털은 타고난, 선천적인 (The big hair on Jenny's face is congenital, inborn.)
fictitious	픽!, 튀셨수 (pick!, ran away)	허구의 보물선 사업으로 돈을 끌어모은 뒤 픽! 튀셨수 (After gathering money with a fictitious treasure ship business, pick! he ran away.)
inebriate	안, 이불이, 에잇! (in, blanket, damn!)	너무 취해서 술집 안에서 "이불이 어딨지? 에잇! 그냥 바닥에서 자자" 하는 술꾼 (A drunkard so inebriated that in the bar he says "Where's the blanket? Damn! Let's just sleep on the floor.")

insurrection	안에, 서, 액션 (in, stand, action)	안에 누워 있는 자들이여, 일어서 액션을 취합시다!하며 폭동을 일으키다 (Those lying inside, let's stand up and take action! starting an insurrection.)
intransigent	안, 넘어오다, 전투 (in, cross into, battle)	국경 안으로 넘어와 전투할 만큼 비타협적인 (Intransigent enough to cross into the border and battle.)
inveterate	안, 뱉어 (in, spit)	입 안에 침을 자꾸 뱉는 습관이 뿌리 깊은 (Having a deep-rooted habit of constantly spitting saliva from the mouth.)
mayhem	매인, 햄 (tied to, ham)	줄에 매인 햄을 서로 먹으려는 대혼란 (Mayhem of everyone trying to eat the ham tied to a string.)
peccable	팩, 꺼, 불 (pack, turn off, fire)	한석봉의 어머니가 팩! 하고 불을 꺼 한석봉이 붓글씨를 쓸 때 과오를 범하기 쉬운 (Han Seokbong's mother goes "pack!" and turns off the fire, making Han Seokbong peccable when writing calligraphy.)
provisional	프로그램, 비잖아 (program, it is empty)	방송 사고로 지금 내보낼 TV 프로가 비잖아, 임시의 방송이라도 내보내! (Due to a broadcast accident, there's no TV program to air now, so broadcast something provisional!)
reimburse	다시, 안, 버스 (again, into, bus)	버스비를 안 내고 내려서 다시 버스 안으로 들어가 버스비를 갚다, 상환하다 (After getting off without paying the bus fare, going back into the bus to reimburse the bus fare.)
squander	습관, 더 (habit, more)	돈을 필요한 양보다 더 쓰는 습관, 즉 낭비하다 (The habit of spending more money than needed — to squander.)

Table 15: KSS Keyword and Verbal Cue

D.3.3 OGR Keyword and Verbal Cue

English Word	**OGR Keyword**	**OGR Verbal Cue**
albeit	얼, 비 (freeze, rain)	비록 ~이기는 하나 그는 얼어붙은 길을 걸으며 비를 맞았다. (Although , he walked on the frozen road in the rain.)
annihilate	안, 아이, 라이트 (inside, child, light)	안에서 아이들이 놀던 건물이 붕괴하자 라이트가 깜빡였다. (When the building where children were playing inside collapsed, the light flickered.)
canny	케냐 (Kenya)	케냐에서 영리한 사업 확장. (Clever business expansion in Kenya.)

esoteric	에서, 테이크 (from, take)	식당에서 비법을 이어받은 요리를 테이크아웃했다. (Took out food that inherited secret recipes from the restaurant.)
incumbent	인, 금, 반지 (person, gold, ring)	재임 중인 그는 인과 금 반지를 받았다. (The incumbent received a person and gold ring.)
insolvent	인, 솔, 벤트 (person, Sol, bent)	인생에서 솔직함을 추구하다 벤처 사업으로 파산한. (Bankrupt from venture business while pursuing honesty in life.)
meddlesome	메달, 섬 (medal, island)	참견하기를 좋아하는 그녀는 메달 수여식에 섬까지 갔다. (She who likes to meddle went to the island for the medal ceremony.)
probation	포, 배, 신 (four, ship, god)	포트폴리오를 배포할 수습 기자가 신문사에서 일했다. (A probationary reporter who would distribute portfolios worked at the newspaper company.)
reckon	레고, 큰 (LEGO, big)	레고로 큰 성을 만들 수 있다고 생각했다. (Thought he could build a big castle with LEGO.)
refurbish	리, 버스 (Li, bus)	리 작업실을 쇄신하여 버스 정류장에서 보는 예술 공간으로 만들었다. (Renovated Li’s workshop into an art space visible from the bus stop.)
resuscitate	리, 서, 시, 테이프 (Li, stand, hour, tape)	리 박사는 환자를 소생시키려 서둘러 시계를 보며 테이프를 확인했다. (Dr. Li hurriedly checked the tape while looking at the clock to resuscitate the patient.)
upheaval	업, 이불 (industry, blanket)	경제적 업 속에서도 이불 속 불안은 대변동의 징조였다. (Even amid economic industry, the anxiety under the blanket was a sign of great upheaval.)
anachronism	아, 낙지, 룬, 이슴 (ah, octopus, rune, -ism)	아, 낙지를 룬으로 변신시키려다 시대착오로 끝났다. (Ah, trying to transform octopus into runes ended in anachronism.)
bureaucracy	비, 오, 러, 크시 (rain, come, Rue, -ksi)	비가 오면 러시아워처럼 되는 관료정치. (Bureaucracy that becomes like rush hour when it rains.)
delirium	딜러, 이음 (dealer, joint)	딜러가 이음 없이 헛소리를 했다. (The dealer spoke deliriously without pause.)
demolish	디, 머리, 시 (D, head, city)	건물을 파괴하여 디테일을 무시하고 머리 속 시각을 그린다. (Demolish buildings, ignore details, and draw the vision in one’s head.)
felon	펠, 언니 (Pel, sister)	펠과 그의 언니와 함께한 흉악범의 계획. (The felon’s plan with Pel and his sister.)

incarcerate	인형, 칼, 새 (doll, knife, bird)	인형과 칼을 가진 새를 지키려다 투옥되고 만다. (Ended up imprisoned trying to protect the bird with a doll and knife.)
ingenious	인기, 뉴스 (popularity, news)	인기를 끌며 뉴스에 나온 기발한 아이디어. (Ingenious idea that gained popularity and appeared on the news.)
recidivist	리, 시, 디, 피스트 (Li, hour, D, fist)	리 마을의 상습범이 시계탑 근처에서 디자이너 가방을 훔치고 피스트를 벌였다. (The recidivist from Li village stole a designer bag near the clock tower and had a fist fight.)
redoubtable	래, 다운, 더블 (Rae, down, double)	무시무시한 래퍼는 다운된 무대에서 더블타임 랩을 했다. (The formidable rapper performed double-time rap on the downed stage.)
remunerate	리, 무, 내다 (Li, nothing, give out)	리 씨는 무더위 속의 노고를 보상했다. (Mr. Li compensated for the hard work in the sweltering heat.)
render	랜, 돌 (LAN, stone)	랜을 돌로 던져 마법을 만들다. (Throw the LAN with stone to create magic.)
repercussion	리본, 커피, 션 (ribbon, coffee, Shawn)	리본을 달고 커피를 마신 션의 하루에 긍정적인 영향. (Positive impact on Shawn's day wearing a ribbon and drinking coffee.)
autopsy	옷, 합시다 (clothes, let's do)	사체의 옷을 확인하고 부검을 합시다. (Let's check the corpse's clothes and perform an autopsy.)
congenital	컨, 제니, 탈 (Con, Jenny, mask)	컨과 제니는 탈을 쓰고 타고난 무대를 빛냈다. (Con and Jenny wore masks and shone on their natural stage.)
fictitious	피크, 티셔츠 (peak, T-shirt)	피크닉 티셔츠에 허구의 이야기를 담았다. (Put fictitious stories on the picnic T-shirt.)
inebriate	인, 애비, 에이 (person, father, A)	술꾼 인영은 애비와 에이급 바에서 시간을 보냈다. (Drunkard Inyoung spent time with father at an A-grade bar.)
insurrection	인, 수레, 션 (person, cart, Shawn)	폭동 중 인파 속에서 수레를 밀고 션이 이끌었다. (During the insurrection, Shawn led by pushing a cart through the crowd.)
intransigent	인, 트랜스, 젠트 (person, trance, gent)	인적이 드문 길에서 비협조적인 그는, 트랜스 음악에 젠트하게 반응했다. (The intransigent person on the deserted road reacted gently to trance music.)
inveterate	인, 배터리 (person, battery)	인내심이 뿌리 깊은 철수는 배터리가 소모될 때까지 참았다. (Cheolsu, with deep-rooted patience, endured until the battery was depleted.)
mayhem	메기, 힘 (catfish, strength)	메기가 힘을 쓰자 대혼란이 일어났다. (When the catfish used its strength, mayhem broke out.)

peccable	배, 꺼, 불 (ship, turn off, fire)	과오를 범하기 쉬운 그는 배를 타고 가다가 전등을 꺼서 불이 꺼졌다. (He who was prone to error turned off the light while on the ship, and the fire went out.)
provisional	프로, 비서, 널 (pro, secretary, you)	프로 프로젝트에서 비서로 임시의 역할을 했다. (Played a provisional role as secretary in the pro project.)
reimburse	림, 버스 (rim, bus)	림에서 버스를 기다리며 돈을 갚다. (Repay money while waiting for the bus at the rim.)
squander	숯, 권, 더 (charcoal, volume, more)	숯을 권 단위로 더 사서 낭비했다. (Wasted by buying more charcoal by the volume.)

Table 16: OGR Keyword and Verbal Cue

D.3.4 PHT Keyword and Verbal Cue

English Word	PHT Keyword	PHT Verbal Cue
albeit	올, 바이트 (all, byte)	비록 올바른 바이트 이기는 하나 그는 망설였다. (Although it was the correct byte, he hesitated.)
annihilate	얼결, 레이더 (accidentally, radar)	얼결에 레이더 시스템이 붕괴했다. (The radar system accidentally collapsed.)
canny	케어, 니은 (care, nieun)	그는 케어를 받으며 니은을 그릴 때도 영리했다. (He was clever even when drawing ’nieun’ while receiving care.)
esoteric	애, 스프링 (child, spring)	그 애는 스프링의 작동 원리를 소수만 이해하는 전문가였다. (That child was an expert in spring operation principles understood by only a few.)
incumbent	인어, 콘센트 (mermaid, outlet)	인어가 콘센트 옆에서 재임 중인 느낌으로 노래했다. (The mermaid sang with a feeling of being incumbent next to the outlet.)
insolvent	인어, 선발대 (mermaid, advance team)	인어가 선발대를 따라가다 파산했다. (The mermaid went bankrupt while following the advance team.)
meddlesome	매달다, 섬 (hang, island)	그는 섬을 매달고 싶다며 모든 일에 참견하기를 좋아했다. (He liked to meddle in everything, saying he wanted to hang the island.)
probation	프로, 봉변 (pro, trouble)	그는 프로처럼 봉변을 수습했다. (He handled the trouble like a pro.)
reckon	레게, 컨셉 (reggae, concept)	그는 레게 스타일이 멋진 컨셉이라고 생각했다. (He thought reggae style was a cool concept.)

refurbish	리포터, 쉬 (reporter, rest)	리포터는 쉬지 않고 방을 쇄신했다. (The reporter renovated the room without rest.)
resuscitate	리더, 스케이트 (leader, skate)	리더는 스케이트를 통해 팀의 사기를 소생시켰다. (The leader revived the team’s morale through skating.)
upheaval	업, 바이블 (industry, bible)	회사의 업계 바이블이 격변을 맞았다. (The company’s industry bible faced upheaval.)
anachronism	어느, 크리스천 (any, Christian)	어느 시대에나 크리스천을 시대착오적인 사람(관습생각)이라 부르는 경우가 있다. (In any era, there are cases where Christians are called anachronistic people.)
bureaucracy	병따개, 러시아 (bottle opener, Russia)	병따개를 들고 러시아에 간 관료. (A bureaucrat who went to Russia with a bottle opener.)
delirium	달리아, 라임 (dahlia, lime)	달리아 꽃밭에서 라임을 곁들이며 헛소리를 늘어놓았다. (He rambled deliriously in the dahlia garden with lime.)
demolish	대물림, 쉬 (inheritance, rest)	대물림된 집을 쉬지 않고 철거했다. (He demolished the inherited house without rest.)
felon	펠레우스, 런던 (Peleus, London)	펠레우스를 런던으로 데려간 흉악범. (The felon who took Peleus to London.)
incarcerate	인가, 샐러드 (approval, salad)	그는 인가를 받지 못해 샐러드를 훔치다 투옥되었다. (He was imprisoned for stealing salad because he couldn’t get approval.)
ingenious	인지, 뉴스 (cognition, news)	그의 인지가 뉴스에 나올 만큼 (생각 등이) 기발했다. (His cognition was ingenious enough to make the news.)
recidivist	리시버, 비슷이 (receiver, similarly)	리시버를 통해 비슷이 행동하는 상습범이 있었다. (There was a recidivist who acted similarly through the receiver.)
redoubtable	라디오, 토플 (radio, TOEFL)	라디오에서 들려오는 무시무시한 뉴스가 토플 준비에 방해가 되었다. (The formidable news from the radio interfered with TOEFL preparation.)
remunerate	라면, 레이더 (ramen, radar)	그는 라면을 먹고 레이더를 고치면 보상하겠다고 말했다. (He said he would compensate if they ate ramen and fixed the radar.)
render	랜드, 더 (land, more)	그는 랜드를 더 재미있게 만들었다. (He made the land more interesting.)
repercussion	리그, 파티션 (league, partition)	리그가 파티션에 미친 영향은 절대적이었다. (The league’s impact on the partition was absolute.)

autopsy	오, 텃세 (oh, territorial behavior)	오랜만에 텃세를 부리는 이웃이 (사체의) 부검 이야기로 사람들을 놀라게 했다. (The neighbor showing territorial behavior after a long time surprised people with autopsy stories.)
congenital	컨테이너, 털 (container, fur)	그는 컨테이너를 열며 털을 정리하는 솜씨가 타고났다. (He had a natural talent for organizing fur while opening containers.)
fictitious	픽, 투표소 (pick, polling station)	픽을 던진 후 투표소에서 모든 이야기를 지어냈다. (After throwing the pick, he made up all the stories at the polling station.)
inebriate	일없이, 리야드 (idly, Riyadh)	그는 일없이 리야드에서 모든 사람을 취하게 했다. (He idly made everyone drunk in Riyadh.)
insurrection	인어, 선입견 (mermaid, prejudice)	인어는 선입견에 맞서 반란을 일으켰다. (The mermaid rebelled against prejudice.)
intransigent	인터넷, 지진대 (internet, seismic zone)	그는 인터넷에서 지진대 정보를 찾으면서도 비타협적인 태도를 유지했다. (He maintained an intransigent attitude while searching for seismic zone information on the internet.)
inveterate	인어, 배터리 (mermaid, battery)	인어의 배터리 사용 습관은 뿌리 깊었다. (The mermaid's battery usage habit was deeply rooted.)
mayhem	매입, 힘 (purchase, strength)	매입으로 인해 힘이 생기면서 아수라장이 되었다. (The purchase gave strength and caused mayhem.)
peccable	패, 커플 (faction, couple)	그는 새로운 패를 내놓고 커플 앞에서 과오를 범하기 쉬운 사람임을 드러냈다. (He revealed himself to be someone prone to error in front of the couple while presenting a new faction.)
provisional	패러디, 저널 (parody, journal)	패러디 저널은 일시적인 인기를 끌었다. (The parody journal gained temporary popularity.)
reimburse	레임덕, 스무 (lame duck, twenty)	레임덕 시기에도 그는 스무 번이나 빚을 갚았다. (Even during the lame duck period, he repaid debts twenty times.)
squander	세관, 더 (customs, more)	그는 세관에서 시간이 더 낭비되었다. (He wasted more time at customs.)

Table 17: PHT Keyword and Verbal Cue

D.4 Metrics

D.4.1 LLM-as-a-judge Prompt

The prompt was originally designed in Korean. For reproducibility, we provide both the original and its English translation.

Prompt	당신은 영어 어휘 학습을 평가하는 채점자입니다. *You are a grader evaluating English vocabulary learning.* 정답 의미: 게으른, 나태한 *Correct meaning: lazy, idle* 학습자 응답: 게으른 *Learner response: lazy* 학습자의 응답이 정답 의미와 일치하는지 평가해주세요. *Please evaluate whether the learner’s response matches the correct meaning.* 다른 표현이나 다른 품사로 설명했더라도 의미가 유사하다면 정답으로 인정합니다. *Accept as correct if the meaning is similar, even if expressed differently or in a different part of speech.* 1(정답) 또는 0(오답)으로만 응답하세요. *Respond only with 1 (correct) or 0 (incorrect).*
Response	1

Table 18: Prompts for evaluating correctness of recognition responses.

Prompt	당신은 영어 어휘 학습을 평가하는 채점자입니다. *You are a grader evaluating English vocabulary learning.* 정답 영단어: squander *Correct English word: squander* 학습자 응답: squander *Learner response: squander* 학습자의 응답이 정답과 일치하는지 평가해주세요. 약간의 오타, 대소문자 차이, 복수형/단수형 차이, 품사 차이 등은 허용합니다. *Please evaluate whether the learner’s response matches the correct answer. Minor typos, case differences, plural/singular differences, and part of speech differences are allowed.* 하지만 같은 뜻을 가지는 다른 영단어는 오답으로 판단합니다. *However, different English words with the same meaning are considered incorrect.* 1(정답) 또는 0(오답)으로만 응답하세요. *Respond only with 1 (correct) or 0 (incorrect).*
Response	1

Table 19: Prompts for evaluating correctness of generation responses.

D.5 Case Study

D.5.1 Correctness Comparison by Word

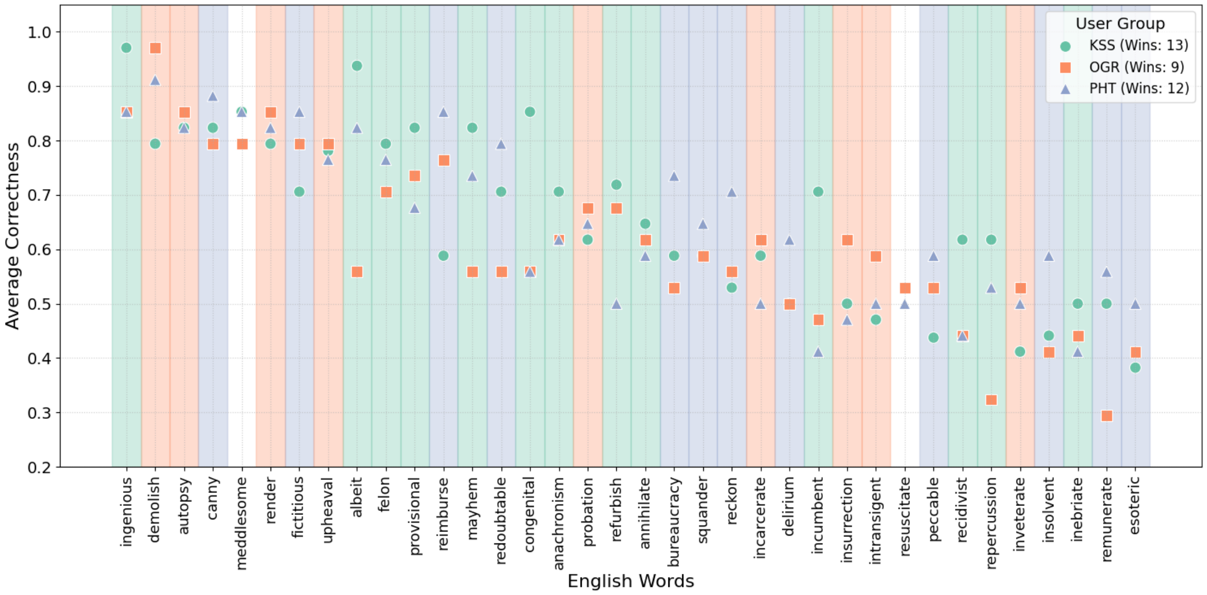


Figure 10: Per-word correctness scores across user groups. Each marker represents the average correctness score for a specific target word, grouped by user condition (KSS, OGR, PHT). For each word, the highest-scoring group is highlighted with a shaded background. The number of wins is comparable between KSS (13) and PHT (12), suggesting no consistent dominance across all words. Given the variation in group performance by word, this pattern motivates further analysis into the word-level characteristics that influence correctness across different groups.

D.5.2 Qualitative Comparison of PHT and KSS

Word (IPA)	**PHT Keyword (IPA)**	**PHT Verbal Cue**	**KSS Keyword (IPA)**	**KSS Verbal Cue**
reckon /ˈrɛkən/	*레게, 컨셉* /l ɛ g ɛ/, /k^{h} ʌ n s ɛ p/	그는 레게 스타일이 멋진 컨셉이라고 생각했다. *He thought reggae style was a cool concept.*	*내 껀* /n ɛ k^{h} ʌ n/	이 금도끼는 내 껀(내 것) 아니라고 생각했다. *He thought this golden axe was not mine.*
render /ˈrɛndər/	*랜드, 더* /l ɛ n d ɨ/, /t ʌ/	그는 랜드를 더 재미있게 만들었다. *He made the land more fun.*	*낸다* /n ɛ n d a/	심부름센터에 돈을 낸다(주다, 제출하다). 그리고 하게 하다. *He pays money to the errand center (to give, to submit), and thus makes something happen.*

Table 20: Examples where PHT outperformed KSS, based on correctness rankings. IPA shown beneath keyword sequence.

Word	PHT Keyword · *POS* · *Similarity*	PHT Verbal Cue	KSS Keyword · *POS* · *Similarity*	KSS Verbal Cue
felon	*펠레우스, 런던* · Noun, Noun · 0.86	펠레우스를 런던으로 데려간 흉악범. *The vicious criminal who took Peleus to London.*	*펠, 놈* · Verb, Noun · 1.15	곤장을 팰 놈, 즉 중죄인, 흉악범. *The one to be beaten with a cudgel—that is, a serious criminal or felon.*
mayhem	*매입, 힘* · Noun, Noun · 1.07	매입으로 인해 힘이 생기면서 아수라장이 되었다. *Because of the purchase, strength arose and chaos ensued.*	*매인, 햄* · Adj, Noun · 1.10	줄에 매인 햄을 서로 먹으려고 수많은 개들이 대혼란, 아수라장. *Countless dogs fought to eat the ham tied to a rope, causing great turmoil and mayhem.*

Table 21: Examples where KSS outperformed PHT, based on correctness rankings. Part-of-speech and phonetic similarity annotated.