

Introducing HYN-KorPus: A Multimodal Neurocognitive Database for Korean

Jaewon Choi, Sumin Jo, Gaeun Lim, Yunju Nam

Hanyang University

Corresponding Author: Yunju Nam (yjnam05@hanyang.ac.kr)

Eye-tracking (ET) and electroencephalography (EEG) are widely used behavioral and neurophysiological measures in reading and sentence processing research. While various language processing databases such as MECO [1] and ZuCo [2] have been developed recently, they remain focused on English and Indo-European languages. Korean was partially included in MECO Wave 1, but no large-scale database for Korean text processing exists.

To address this gap, we present HYN-KorPus (Hanyang Neuro-cognitive KOREan corPUS), the first multimodal (ET + EEG) neurocognitive database for Korean. Using materials that contain linguistic features such as case markers, relative clauses, and conjunctions, ET is recorded during visual presentation of sentences and texts, and EEG during visual presentation of sentences and auditory presentation of both sentences and texts. Data collection is ongoing, with a target of at least 500 participants. Collected data are undergoing preprocessing for release and will be made publicly available via OSF upon completion. HYN-KorPus provides Korean-specific linguistic annotation using the Sejong POS tagger, distinguishing case particles, stems, and endings – features that universal POS tags (UPOS) fail to capture, resulting in contradictory patterns in MECO Korean data [3, 4] (see Appendix). In a pilot analysis of a 70-participant subset of the text-reading (visual) ET condition [4], Korean readers showed decreased reading time at sentence-final predicates, a reversed pattern from the wrap-up effect with increased reading times at sentence endings reported in English [5]. Japanese, another head-final language, likewise showed no clause wrap-up effect [6]. A Korean BERT model (KLUE/BERT-base), by contrast, showed increased attention at sentence-final predicates, exhibiting a pattern closer to the English-style wrap-up effect.

Korean readers showed longer fixation times on unfamiliar proper nouns and technical terms, whereas model attention at these same regions remained low. KLUE-BERT attention correlated negatively with total reading time overall (Figures 1, 2). Corpus-wide analysis confirmed the same mismatch: at sentence endings (EOS), model attention was high while Korean readers' reading time was low (Figure 3).

This corpus provides empirical understanding of processing in SOV and agglutinative languages, serving as a cognitive benchmark for Korean language models, and presents a model for constructing neurocognitive data for under-resourced languages beyond English-centered language processing research.

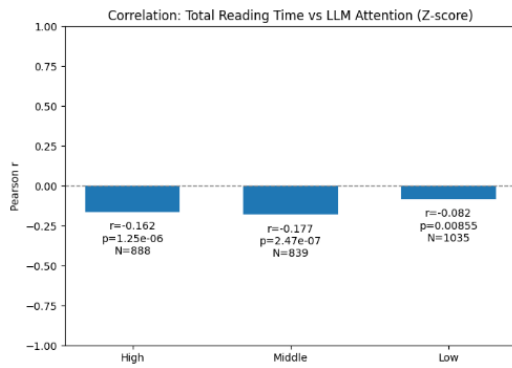


Figure 1: Negative correlations between human total reading time and KLUE-BERT attention across three difficulty levels

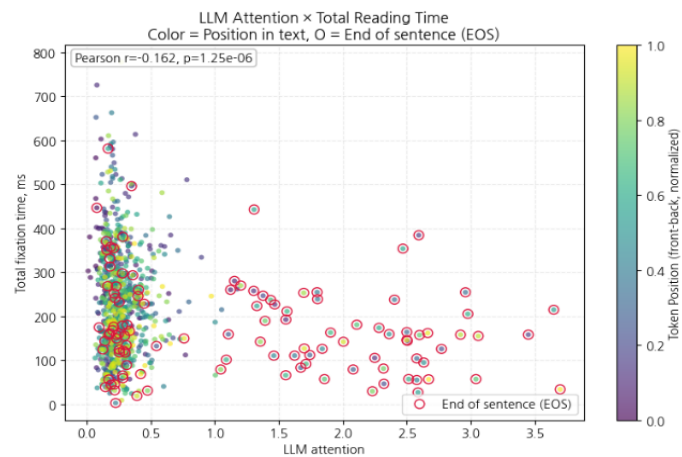


Figure 3: Corpus-wide Analysis of Human Reading Time vs. LLM Attention in Non-fiction Text

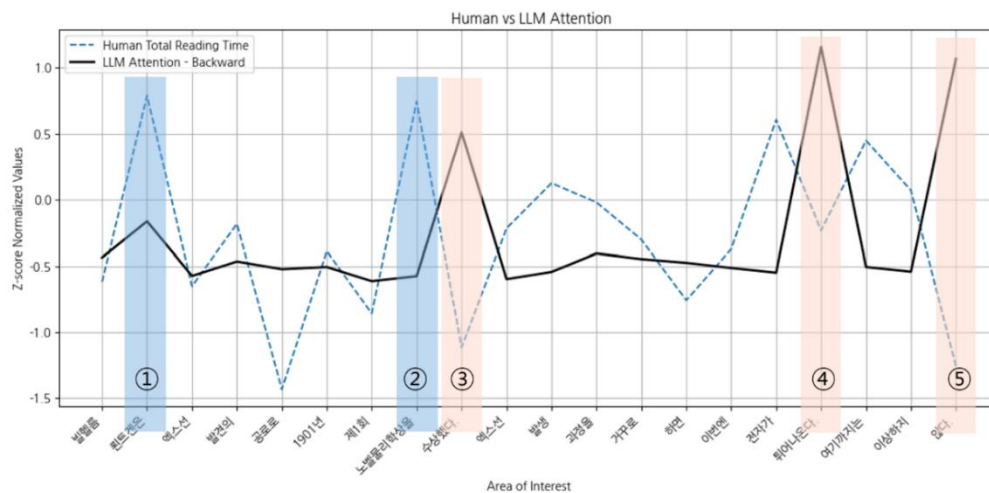


Figure 2¹: Human vs. LLM Attention Patterns in a Sample Sentence

References

- [1] Siegelman, N., Schroeder, S., Acartürk, C., et al. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54, 2843-2863.
- [2] Hollenstein, N., Rotsztein, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1), 180291.
- [3] Brandl, S., & Hollenstein, N. (2022). Every word counts: A multilingual analysis of individual human alignment with model attention. *arXiv preprint*.
- [4] Choi, J. W., Jo, S. M., Lim, G. E., & Nam, Y. J. (2025). Comparing LLM attention and human eye-tracking data in Korean reading. *Presented at the 50th Anniversary Conference of the Linguistic Society of Korea*.
- [5] Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, 53(4), 1061-1080.
- [6] Asahara, M. (2018). Between reading time and clause boundaries in Japanese-wrap-up effect in a head-final language. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

¹ Blue shading indicates areas where human reading time is high but LLM attention is low (proper nouns). Red shading indicates areas where LLM attention is high but human reading time is low (predicates).

① Roentgen-eun (뢴트겐은) – “Roentgen” (unfamiliar proper noun) + topic marker, ② Nobel-mulli-haksang-eul (노벨물리학상을) – “Nobel Prize in Physics” + object marker, ③ susang-haet-da (수상했다) – “received (the prize)” (predicate), ④ twieona-onda (튀어나온다) – “protrudes” (predicate), ⑤ an-ta (않다) – “not” (negative predicate)

Korean Linguistic Features and Tagging Challenges

Korean is an SOV language with agglutinative characteristics, where the grammatical and semantic roles of sentence constituents are not determined solely by word order but are explicitly marked by case particles. Additionally, Korean adjectives, unlike English adjectives, function as predicates that conjugate similarly to verbs. Therefore, morpheme-level annotation including particles and endings, rather than word-level POS tags, is essential for analyzing Korean sentence processing.

1. Limitations of UPOS for Korean

While UPOS (Universal POS tags) facilitates cross-linguistic comparability, its application to Korean data fails to adequately reflect morphological structure and grammatical categories. Case particles and endings are crucial for sentence interpretation in Korean, yet UPOS analysis either oversimplifies them into broad categories such as ADV, ADP, PART, and SCONJ, or classifies them contrary to their actual grammatical functions.

2. UPOS vs. Sejong Tag annotation²

Example 1: 아름다운 해변에 가서(going to the beautiful beach)

	아름다운	해변에	가서
UPOS ³	아름답+ㄴ/ADJ	해변+에/ADV	가+서/SCONJ
Sejong ⁴	아름답/VA-I + ㄴ/ETM	해변/NNG + 에/JKB	가/VV + 어서/EC

Example 2: 저렴한 오렌지주스(cheap orange juice)

	저렴한	오렌지주스
UPOS	저렴+하+ㄴ/VERB	오렌지주스/SCONJ
Sejong	저렴/XR + 하/XSA + ㄴ/ETM	오렌지/NNG + 주스/NNG

3. Why These Differences Matter for Cognitive Processing

Role of Case Particles: In Korean, case particles such as “-가, -을/를, -에게, -에” are core units marking subject, object, and dative cases. To measure processing load of case particles in eye-tracking research, the boundaries between particles and nouns must be clearly distinguished, yet UPOS treats them as integrated units.

Predicate Conjugation: Korean adjectives, unlike English adjectives, are predicates that conjugate similarly to verbs, separating into stems and endings with independent grammatical functions. For example, “아름다운” combines the adjectival stem “아름답-” and the adnominal ending “-ㄴ”. To distinguish neural signals of stem processing (lexical meaning) from ending processing (grammatical function) in EEG research, morpheme-level analysis is essential.

Connective and Adnominal Endings: To analyze Korean clause connection and noun modification mechanisms distinct from English, connective endings (“-어서”) and adnominal endings (“-ㄴ”) must be tagged independently.

For these reasons, HYN-KorPus is designed to directly link Korean morpheme-level processing characteristics with eye movements and EEG responses through Sejong POS tagger-based annotation.

² Examples are from the MECO Wave 1 Korean dataset [1].

³ UPOS Tag: ADJ = Adjective, ADV = Adverb, VERB = Verb, SCONJ = Subordinating Conjunction

⁴ Sejong Tag (parser: Kiwi): A Korean-specific morphological analysis system developed in the 21st Century Sejong Project, which precisely distinguishes Korean-specific grammatical categories including particles, endings, and stems. HYN-KorPus uses the Kiwi parser to apply Sejong tagset annotation. VA-I = Irregular Adjectival Predicate Stem, ETM = Adnominal Ending, NNG = Common Noun, JKB = Adverbial Case Particle, VV = Verbal Predicate Stem, EC = Connective Ending, XR = Root, XSA = Adjectival Derivational Suffix