SoriGraph: A New Database of Visual Feature-Level Descriptions of Written Korean

Wednesday Bushong*, Hala Habahbeh*, Ryan Jiang†, Yoolim Kim†
*Wellesley College, Wellesley, MA, USA
{wb104, hh108}@wellesley.edu
†Carleton College, Northfield, MN, USA
{jiangr, ykim6}@carleton.edu

Abstract

Phoneticians and phonologists have developed featural systems that enable systematic description of human speech sounds. However, no such systems exist for describing the visual features of writing systems. It is critical to understand the features of writing systems given their central role in many language users' everyday experience. Just as phonetic and phonological features provide insight into speech perception, visual features can play a similar role for studying reading. In this paper, we introduce SoriGraph, a database of visual feature descriptions and IPA transcriptions for the full lexicon of Korean, drawing on a recent large-scale study of the visual features of writing systems. This database enables analysis of the visual and phonological properties of Korean and will be a critical resource for researchers. We describe the construction of the database and provide an overview of several potential uses of the database, and demonstrate one potential usage (information-theoretic analysis of lexicon structure).

Keywords: Korean, Hangul, visual features

1. Introduction

1.1. Korean and Hangul

For this project, we focus on Korean and its writing system, Hangul.[1] Typologically, Korean is an agglutinative, head-final language with rich verbal morphology. In terms of its sound system, Korean has altogether nineteen consonants and ten vowels.[2] Hangul is system wherein the individual graphemes represent units of sound in the language. The graphemes are assembled into blocks, as shown in Figure 1, which illustrates two blocks (left) and their constituent graphemes, or *jamo*, assembled in linear and non-linear fashion (right). There are 40 unique *jamo* characters.

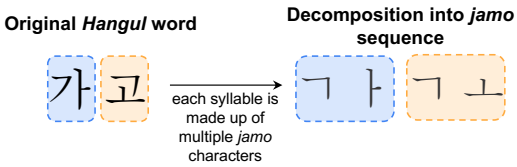

Figure 1: Each Hangul character represents a spoken syllable and is composed of *jamo* characters corresponding to individual (or some biphone sequences of) sounds. This figure shows the original Hangul form of 가고 (Romanized: *gago*) on the left, and its linearized *jamo* sequence on the right.

1.2. Describing the Visual Features of Writing Systems

Phoneticians and phonologists have developed systematic descriptions of the acoustic and articulatory features of speech. Relatively fewer analogous systems exist for describing the visual features of written language, which often focuses on either a limited subset of writing systems (e.g., Primus, 2004), or on specific visual properties (e.g., topology; Changizi and Shimoji, 2005). *Glyph* is a recent study which aimed to develop systematic visual descriptions of 43 writing systems, including Hangul, by crowdsourcing visual feature descriptions from a large number of participants via a gamified task (Kim et al., 2025). Participants' task was to divide the inventory of *jamo* characters into two sets according to a visual rule they specified. Rules were then validated by retesting participants on their own rules. After collecting many such validated rules, Kim et al., 2025 found the minimal set of visual feature rules that allows for efficient and unique identification of each *jamo* character, which resulted in 19 rules. Table 1 shows each visual feature rule, along with the *jamo* characters that satisfy the rule and the original participant description of the visual feature.

Given the crowdsourcing nature of the project, while some visual feature rules map clearly to visual properties, some are more opaque. For example, Rule 1 ("two of something") and Rule 2

[1]We acknowledge the vibrant debate on classifying Hangul as a system; here, we present Hangul generally as a system that encodes the sounds of the language, which we believe is sufficient for the purposes of our present research.

[2]The number of phonemes may vary depending on how semi-vowels and diphthongs are considered, but for our purposes, we consider the phonemic inventory to comprise twenty-nine phonemes.

Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026, pages 45–49
12 May, 2026.

Rule	*jamo* characters	Original participant description
Rule 1	ㄲㄸㅃㅆㅉ	“two of something”
Rule 2	ㅏㅐㅑㅒㅓㅔㅕㅖㅘㅙㅚㅝㅞㅟㅢㅣ	“Allong”
Rule 3	ㅅㅆㅈㅉㅊ	“man symbol”
Rule 4	ㅅㅆㅇㅈㅉㅊㅎㅡㅢㅣ	“does not contain a right angle”
Rule 5	ㄲㄸㅃㅆㅎㅔㅖㅘㅙㅚㅝㅞㅟㅢ	“Buchstaben, die aus mindestens zwei voneinander getrennten Elementen bestehen”
Rule 6	ㅁㅂㅃㅇㅍㅎㅒ	“Characters with enclosed spaces.”
Rule 7	ㅗㅘㅙㅚ	“Right way up boat”
Rule 8	ㅁㅂㅃㅇㅎㅒ	“AT LEAST ONE ENCLOSED SPACE”
Rule 9	ㅑㅒㅕㅖㅛㅠ	“All characters with a line and two touching tiny rectangles”
Rule 10	ㄱㄴㅈㅉㅊㅎㅗㅘㅙㅚㅛㅜㅝㅞㅟㅠㅡㅢ	“one long horizontal line”
Rule 11	ㅒㅖㅔㅖㅙㅞ	“more than one fullsized vertical line”
Rule 12	ㄱㄲㄷㄸㅂㅃㅅㅆㅈㅉ	“Lettres doubles”
Rule 13	ㅏㅑㅓㅔㅕㅖㅗㅘㅛㅜㅠ	“Halbe Leiter nur gerade”
Rule 14	ㄷㅋㅐㅑㅔㅕㅚㅟㅠ	“Contains three straight lines”
Rule 15	ㄱㄲㅋㅓㅔㅕㅖㅙㅞ	“faces left”
Rule 16	ㄱㄴㄷㄹㅂㅡㅣ	“2 ends”
Rule 17	ㅏㅐㅑㅒㅓㅔㅕㅖㅗㅘㅙㅚㅛㅜㅝㅞㅟㅠ	“petit tiret”
Rule 18	ㄴㄷㄸㅍㅣ	“could pass for a roman numeral”
Rule 19	ㄱㄲㅋㅓㅔㅕㅖㅙㅞ	“a crossbar, with stuff only below it”

Table 1: Visual rules for *jamo* characters from the *Glyph* project (Kim et al., 2025).

(“Allong”) have easily described visual correlates: symmetry (Rule 1) and long vertical stroke (Rule 2). But Rule 7 (“Right way up boat”) and Rule 18 (“could pass for a roman numeral”) have no clear basic features and seem instead to draw on holistic combinations of visual features or analogy to other systems.

1.3. The Present Study

The goal of the present project was to construct a comprehensive database of the Korean lexicon that integrates orthographic, phonological, visual feature, and word frequency information into a single unified resource.

This database, which we dub SoriGraph(a combination of the word for sound in Korean ’sori’, and ’graph’), enables analyses of written Korean that were previously impossible—for example, estimating visual similarity between words beyond rough-grained measures like orthographic similarity, or investigating the relationship between the visual features of a *jamo* character and the phonological features of the sound associated with the character. We hope this database will be a useful resource for researchers interested in the visual structure of written Korean, or the relationship between spoken and written Korean. The database is available at the following link: `https://osf.io/fsb6c`. We provide the database in two file formats. One is a .RDS file readable using the programming language R (R Core Team, 2025). We chose this format because one variable in the database contains matrix entries (the visual feature descriptions of full words) which are difficult to represent in ‘flat’ comma- or tab-separated file formats. We also include the database as a comma-separated value (.csv) format excluding the visual feature matrices, along with the original visual rule-to-*jamo* mapping as another .csv file, so that researchers can compute the word feature matrices themselves.

2. Database Creation

The database includes 282,348 unique word forms. Each word contains the following information: (1) its written Hangul form; (2) its IPA transcription; (3) a decomposition of its Hangul form into a linear sequence of *jamo* characters of length n; (4) an $n \times 19$ matrix containing the visual feature encoding of each *jamo* character; and (5) word frequency information (available for a subset of the words). The sections below describe each step of this process in more detail.

2.1. Dictionary and Preprocessing

We exported the entirety of 표준국어대사전 (*Standard Korean Dictionary*), a comprehensive dictionary of standard Korean (국립국어원 [National Institute of the Korean Language]). This resulted in 436,142 entries. The dictionary includes non-word entries such as affixes, particles, interjections, and multiword expressions, which we removed. We

also removed any entries containing non-Hangul orthographic elements (e.g., *hanja*, Chinese characters). The dictionary also includes separate entries for each meaning of homonyms and polysemous words; we collapsed these distinctions so that every unique written form corresponded to only one entry. We then stripped any special characters from the word form representations (e.g., some entries contained dashes marking morpheme boundaries or parentheticals with further indexing information). This left us with 282,348 unique entries in the database.

2.2. IPA Transcriptions

For IPA transcription, we utilized the Cross-Linguistic Phonological Frequencies (XPF) corpus (Cohen Priva et al., 2021), a resource that provides standardized orthography-to-IPA transcriptions for a wide variety of languages, including Korean. Our procedure began by extracting and identifying all unique syllables (i.e., Hangul characters) present in our database and retrieving their corresponding phonemic transcriptions from the XPF corpus. To avoid issues in rendering across devices, we converted each IPA character to its Unicode value (e.g., ŋ is represented as "U+014B"). Complex IPA characters are represented as a concatenated string of their component Unicode characters (e.g., t^h is represented as "U+0074U+02B0"). For each word entry in the database, we use these mappings to construct a character string of Unicode IPA characters (separated by spaces). One limitation to be aware of is that the XPF corpus does not account for systematic sound changes that cross syllable boundaries, instead assuming a fixed syllable-to-IPA mapping. Thus, not all word transcriptions are an accurate reflection of their spoken realization.

2.3. Hangul-*Jamo* Decomposition

As described above (section 1.1), Hangul is an writing system in which individual phonemic units, including consonants, vowels, and certain biphone or diphthong sequences, are represented by *jamo* characters that combine to form orthographic syllable blocks (Taylor, 1980). Because the syllable block serves as the primary orthographic unit, while the information necessary for visual feature analysis resides at the *jamo* level, we decomposed each word form in our dataset into its constituent syllables and then further segmented them into a linear *jamo* character sequence using the `unicodedata` library in Python.[3] Like the IPA transcriptions, we represented each *jamo* character according to its Unicode representation (e.g., ㄱ [Romanized as *g*] corresponds to "U+1100").

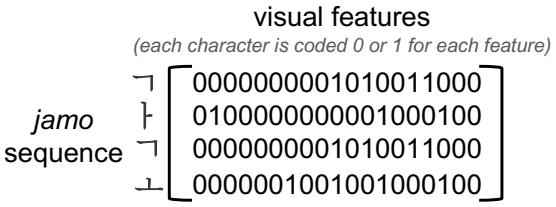

Figure 2: We describe each lexical entry using a matrix with the rows representing each *jamo* character in sequence, and the columns with the binary visual feature vector describing the *jamo* character.

2.4. Visual Feature Matrices

We represent each *jamo* character as a binary vector of length 19. These correspond to the 19 visual feature rules from the *Glyph* project of Kim et al. 2025, that allow for each *jamo* character to be uniquely identifiable.[4] A value of `0` indicates that the character does not satisfy the feature (and vice versa for a value of `1`). We repeat this process for every *jamo* character in the word, resulting in an $n \times 19$ matrix, where *n* is the word length in *jamo* characters. Figure 2 shows an example for the word 가고 (Romanized: *gago*).

2.5. Word Frequency

We obtained frequency measures from an analysis of the Sejong National Corpus (ELRA) conducted by Lee et al. 2016. While Lee et al. 2016 contains a list of only the 5,000 most frequent Korean words, the authors generously provided us with comprehensive frequency estimates of all words occurring in the corpus at least once. The Sejong National Corpus contains both spoken and written language; the frequency estimates Lee et al. 2016 provide are (1) the total number of occurrences of the word form in the corpus, and (2) the average of the spoken and written frequencies of the word form (per 100,000 words). Since our database does not include separate entries for homonyms or polysemous words, we sum the frequencies of all entries associated with each unique word form. This resulted a total of 145,454 entries (51.5%) in our database containing word frequency information.

[3] Unicode metadata includes the linear decomposition of *jamo* characters which make up each unique Hangul character.

[4] That is, we did not identify visual features from the raw data of Kim et al. 2025, but used their identified set of 19 visual feature rules for each *jamo* character.

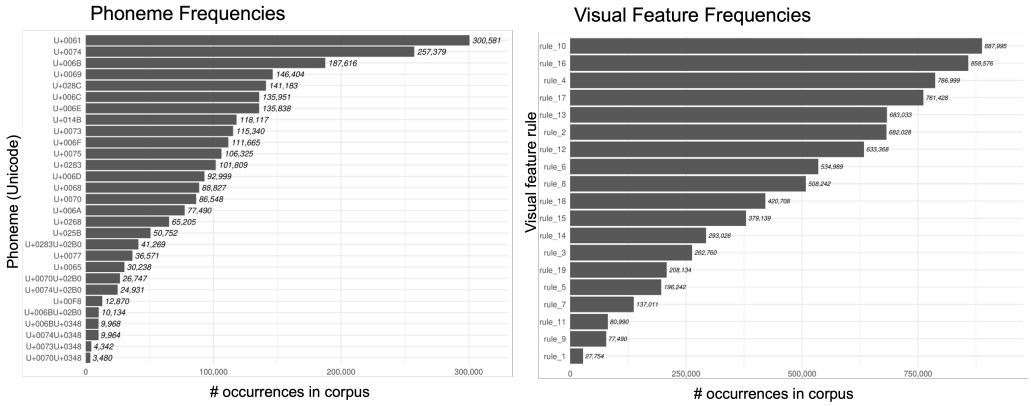

Figure 3: Raw frequency of occurrence of each phoneme (left panel) and visual feature (right panel) in the database.

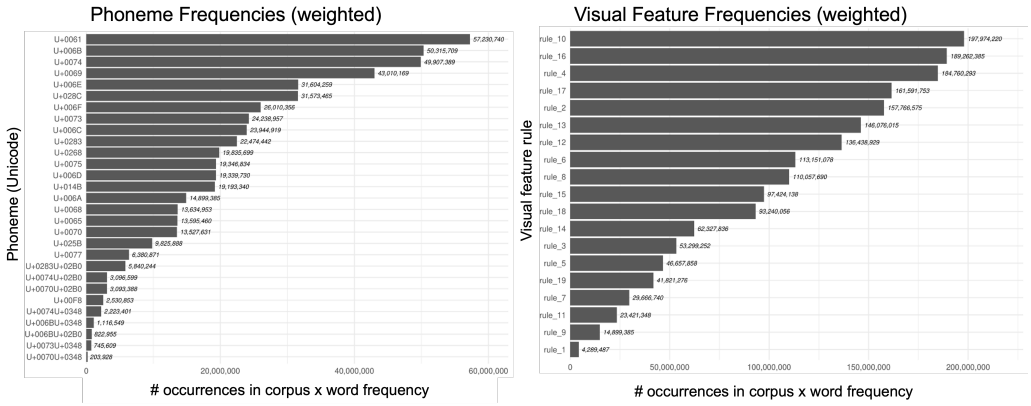

Figure 4: Frequency of occurrence, weighted by word frequency, of each phoneme (left panel) and visual feature (right panel) in the database.

3. Possible Uses of the Database

There are many ways SoriGraph can be utilized by researchers. We briefly describe two possibilities below.

3.1. Information-Theoretic Analysis of Language Structure

Cross-linguistically, the units of spoken language tend to follow similar distributions. For example, word frequencies follow a Zipfian distribution, reducing the entropy of the lexicon and allowing for efficient communication (Zipf, 1949). Phoneme frequencies also follow closely related distributions (Macklin-Cordes and Round, 2020; Tambovtsev and Martindale, 2007).

A natural question, then, is whether corresponding written language units follow a similar pattern. Korean provides a unique opportunity for testing this question, given that the explicit design intent of Hangul was to mirror the spoken language (Sohn, 1999). Is this reflected in the distribution of visual features? We conducted a simple analysis to begin to test this question. Figure 3 shows the frequency of each phoneme (left panel) and visual feature (right panel) across our database, and Figure 4 shows these values weighted by the frequency of the words they occur in. Upon first inspection, while the phonemes of the language appear to show a characteristic (sub-)Zipfian distribution, the visual features seem to be distributed differently (the relationship between frequency and frequency rank is closer to linear rather than Zipfian). This is not meant to be taken as a strong claim, but rather a demonstration that such questions can be asked using SoriGraph. Using SoriGraph, future research can investigate this relationship further and explore its potential implications for written language production and processing.

The database will also allow for other kinds of information-theoretic comparisons between the written and spoken language; for example, examination of which visual features vs. phonemes carry the greatest functional load in the lexicon, and whether these factors may influence diachronic changes.

3.2. Stimulus Design for Reading Experiments

Orthography is known to be an important factor in reading. Access to the visual feature structure of words will allow researchers to compute new measures that can be used to design future reading experiments. For example, our database can support the development of new measures of visual feature neighborhood that extend beyond relatively rough-grained measures like orthographic neighborhood density (OND).

4. Bibliographical References

Mark A. Changizi and Shinsuke Shimoji. 2005. Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B: Biological Sciences*, 272:267–275.

Uriel Cohen Priva, Emily Strand, Shiying Yang, William Mizgerd, Abigail Creighton, Justin Bai, Rebecca Mathew, Allison Shao, Jordan Schuster, and Daniela Wiepert. 2021. *The Cross-linguistic Phonological Frequencies (XPF) Corpus manual*. Accessible online, https://cohenpr-xpf.github.io/XPF/manual/xpf_manual.pdf.

ELRA. Sejong Corpus. http://universal.elra.info/product_info.php?cPath=42_43&products_id=1975.

Yoolim Kim, Marc Allassonnière-Tang, Helena Miton, and Olivier Morin. 2025. The phonology of letter shapes: Feature economy and informativeness in 43 writing systems. *Journal of Memory and Language*, 142:104620.

Sun-Hee Lee, Seok Bae Jang, and Sang Kyu Seo. 2016. A frequency dictionary of Korean: Core vocabulary for learners.

Jayden L. Macklin-Cordes and Erich R. Round. 2020. Re-evaluating phoneme frequencies. *Frontiers in Psychology*, 11.

Beatrice Primus. 2004. A featural analysis of the modern roman alphabet. *Written Language Literacy*, 7:235–274.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ho-min Sohn. 1999. *The Korean language*. Cambridge University Press.

Yuri Tambovtsev and Colin Martindale. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, 4(2):1–11.

Insup Taylor. 1980. *The Korean writing system: An alphabet? A syllabary? a logography?*, page 67–82. Springer US.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort.

국립국어원 [National Institute of the Korean Language]. 표준국어대사전 [Standard Korean language dictionary]. https://stdict.korean.go.kr.