A Mutual Information-based Approach to Quantifying Logography in Japanese and Sumerian

Noah Hermalin
Department of Linguistics
University of California, Berkeley
Berkeley, CA 94720 USA
nmhermalin@berkeley.edu

Abstract

Writing systems have traditionally been classified by whether they prioritize encoding phonological information (**phonographic**) versus morphological or semantic information (**logographic**). Recent work has broached the question of how membership in these categories can be quantified. We aim to contribute to this line of research by treating a definition of logography which directly incorporates morphological identity. Our methods compare mutual information between graphic forms and phonological forms and between graphic forms and morphological identity. We report on preliminary results here for two case studies, written Sumerian and written Japanese. The results suggest that our methods present a promising means of classifying the degree to which a writing system is logographic or phonographic.

1 Introduction

Writing systems vary regarding how much they encode phonological versus morphological (or semantic) information: systems which prioritize conveying phonological information are conventionally labeled as **phonographic**, and systems which prioritize morphological/lexical information are labeled as **logographic** (Daniels and Bright, 1996; Joyce and Borgwaldt, 2013). While this taxonomic split is convenient as a broad-strokes shorthand for classifying different writing systems, more precise categories, as well as more precise definitions for existing categories, remain a point of ongoing research, and debate, among scholars of writing systems.

Defining what it means to be logographic has been a particular point of inconsistency in the literature on writing systems; see section 2 of Sproat and Gutkin (2021) for a review. While cases where a single character maps to an entire word would generally be considered logographic and cases where a single character always maps to a particular segment would be considered phonographic, the numerous in-between cases present points of possible contention. For example, the fact that English spells many homophones differently (e.g., *where, wear, ware*), while spelling distinct allomorphs of a given morpheme the same way (e.g., the root in ***heal*** and ***heal****th*), has motivated treating written English as somewhat logographic. (Sproat, 2000; Rogers, 2005; Sproat and Gutkin, 2021).

Relative to any particular category definitions is the question of how one might quantify the degree to which a writing system belongs to that taxonomic category. While not always framed as a matter of typology, a few methods have been used to quantify the consistency of character string-sound string mappings (orthographic transparency/depth). These include using binary consistent/inconsistent distinctions (e.g., Ziegler et al. (1996, 1997)), entropy-based measures (e.g. Treiman et al. (1995); Borgwaldt et al. (2004); Protopapas and Vlahou (2009); Siegelman et al. (2020)), and machine learning (Marjou, 2019; Rosati, 2022). With regards to consistency of mapping, results differ depending on whether one considers the perspective of the reader or the writer: it's possible to have ambiguity of reading but not spelling (e.g., English *bass*), or vice versa (e.g. English /ɹa͡it/). Penn and Choma (2006) and Sproat and Gutkin (2021) more directly focused on the question of quantifying taxonomic category membership (rather than orthographic transparency). We aim to build on the work in this vein by proposing a simple metric of logography which incorporates graphic forms, phonological forms, and morphological identity.

1.1 The Study of Sproat and Gutkin (2021)

Recently, the question of quantifying category membership has been directed at measuring how logographic a system is. Responding to an earlier attempt by Penn and Choma (2006), Sproat and Gutkin (Sproat and Gutkin, 2021) propose a hand-

Proceedings of the Workshop on Computation and Written Language (CAWL 2023), pages 105–110

ful of ways by which one might quantify degree of logography, relative to a working definition of logography.

Sproat and Gutkin discuss two different treatments of logography. By their **distinct homophones** notion of logography, a system is more logographic if identical phonological forms are not necessarily spelled the same. The rationale behind the distinct homophones approach is: given that a maximally phonographic system will spell all words based solely on their phonological form (thus not discriminating between homophones), deviation from this ideal constitutes a higher degree of logography. Their **uniform spelling** treatment of logography treats a system as more logographic if the same morpheme is always spelled the same (despite surface variation). Citing convenience and availability of reliable data, Sproat and Gutkin use their **distinct homophones** definition for their studies, though they note that the **uniform spelling** approach is also valid.

Sproat and Gutkin compare three different classes of methods for quantifying logography. Their *S* measure is based on the attention mechanism of an RNN that maps phoneme strings to written character strings, in context. A higher *S* means that a system is more logographic, since more attention needs to be given to surrounding context in order to know how to spell a word. As a simple baseline for comparison, their lexical *L* measure computes the average number of spellings *s* per pronunciation *p* (drawn from a dictionary *D* or corpus *C* of *p* types/tokens). A higher *L* means a system is more logographic.

$$L_{type} = \frac{1}{|D|} \sum_{p \in D} |s(p)| \quad \text{and} \quad L_{token} = \frac{1}{|C|} \sum_{p \in C} c(p)|s(p)| \tag{1}$$

Their *E* measure is based on uncertainty of spelling given pronunciation. For their type-based analyses, this was done as the mutual information between written forms $\mathcal{W}$ and pronunciations $\mathcal{P}$:

$$E_{type} = H(\mathcal{W}) - H(\mathcal{W}|\mathcal{P}) \tag{2}$$

The data for Sproat and Gutkin's main experiment was the Bible in 9 languages: English, Hebrew, French, Russian, Swedish, Finnish, Korean, Chinese (at the character- and word-level), and Japanese (Christodouloupoulos and Steedman, 2015); they also ran studies using data from Wikipedia for Finnish, Japanese, English, and Korean, as well as on the Bible again for additional languages. These languages' writing systems range from more (Chinese, Japanese) to less (Finnish, Korean) logographic, in terms of how they are conventionally treated by scholars of writing systems and of those languages. It should be noted that the pronunciations for their main data were automatically generated from their target texts, and their pronunciation generators did not have any homograph disambiguation (a factor which they acknowledge as a shortcoming, and address to some extent in their section 6.5).

Sproat and Gutkin's attention-based *S* measures most closely aligned with how logographic their target writing systems are generally considered to be: *S* scores were lowest for Finnish, Swedish, and Korean, and were highest for Japanese and Chinese. Their *L* measures were less reliable, but still somewhat consistent with expectations. While their E_{type} performed decently, it had a few unexpected outcomes, such as ranking character-level Chinese as too phonographic and Swedish as too logographic, leading Sproat and Gutkin to favor their *S* measure. However, it may be that these *E* results were a consequence of only considering pronunciations and spellings, without also incorporating morphology. A mutual information approach which also includes morphological information still has the potential to match, or outperform, their *S* measure.

2 Logography via Morpheme Identity

One can view graphic forms and spoken forms as two points of a triangle, with the third point being semantics or lexical identity[1]. Since Sproat and Gutkin (2021) focus on their **distinct homophones** interpretation of logography, the role of morphology (and semantics) do not play a role in their experiments. We aim to add morpheme identity to the mix, which helps to complete the missing side of the triangle. Given that traditional definitions of logography have placed emphasis on the role of morphology or semantics in how words are read and written, a complete account of how to quantify logography would benefit from including the graphic form-morphology leg of the triangle.

[1]This kind of 'triangle' model was popularized by Seidenberg and McClelland (Seidenberg and McClelland, 1989). While it's true that that work was focused on modelling how humans read, not on writing system typology, this kind of triangle schematic nonetheless serves as a useful representation of how the components of spoken/written language can connect to each other, which is relevant for classifying writing systems.

Intuitively, if a system is more logographic, then graphic forms and morphological identity will provide more information about each other; if a system is more phonographic, then graphic forms and phonological forms will provide more information about each other. We propose that the degree to which a writing system is logographic l can be quantified by comparing the mutual information between graphic forms G and morphemes M with the mutual information between graphic forms G and phonological forms P:

$$l = I(G; M) - I(G; P) \quad (3)$$

One advantage that mutual information has in this context is its symmetry, with the consequence that it's agnostic as to whether the writing system is being analyzed from the reader's perspective or the writer's perspective. Measures such as Sproat and Gutkin's S and L metrics can only be done from one of the two reader/writer perspectives at a time (Sproat and Gutkin's experiments only took the writer's perspective); while there's nothing stopping one from getting those measures from both perspectives separately, this would give two separate measures rather than the single unified measure that mutual information offers.

3 Data - Sumerian and Japanese

As an initial testing ground, we focus on the writing systems of two unrelated languages: Modern Japanese and Ur III Sumerian. These writing systems are considered to be highly logographic, with morphemes often being spelled with a single character. Sumerian and Japanese happen to both be agglutinating in their morphology, with additional similarities including postnominal case marking, root+affix verbal morphology, and extensive use of compounding (Shibatani, 1990; Michalowski, 2004); given the role that morphology plays in our analysis, it is convenient to start by considering two languages which, by chance, have similar morphological profiles and similar writing systems. In addition, this choice of writing systems also allowed us to have one writing system (Japanese) which was considered by Sproat and Gutkin, and one writing system (Sumerian) which hasn't yet been explored in this vein. To our knowledge, this work constitutes the first attempt at quantifying how logographic or phonographic written Sumerian is. These systems were also chosen in part because of data availability and author background.

To avoid any diachronic variation within Sumerian, all of the Sumerian data were from documents composed during the Ur III period (c. 2112-2004 BC), drawn from 71,712 Ur III administrative documents within ORACC, the Open Richly Annotated Cuneiform Corpus (Tinney and Robson, 2014). This corpus was chosen for its robust size and lexical and morphological annotation. During cleaning, we removed tokens which contained damaged written forms, had uncertain readings/translations, or were proper nouns. Morphologically complex tokens, including compounds and inflected forms, were further processed into morpheme-length (rather than word-length) tokens via both automated and manual parsing by the first author. This resulted in a total of 1,875,351 morpheme-sized tokens.

Japanese data were drawn from BCCWJ, the Balanced Contemporary Corpus of Written Japanese (Maekawa et al., 2014). BCCWJ tokenizes by lemma, with each token annotated with graphic form and phonological form information. Morphologically complex tokens were further processed into morpheme-length tokens by the first author. The analyses reported here include the 5,000 most frequent lemmas, with a total of 62,929,634 morpheme tokens.

3.1 Morphological Parsing

BCCWJ and ORACC both tokenize at the word rather than the morpheme level. As such, some additional processing was needed to get morpheme-sized tokens out of morphologically complex words, particularly compounds and words with grammatical affixes. For Sumerian compounds (548 types), morphological parsing was done completely manually by the first author based on background knowledge and use of the electronic Pennsylvania Sumerian Dictionary (ePSD2)[2]; some parses were also run by someone more knowledgeable on Sumerian to double check their validity. Parsing grammatical morphology on nouns and verbs and aligning the parses with characters was handled automatically by a Python script written by the first author. Because of the complexities of Sumerian verbal morphology, automating exactly which characters mapped to which morphemes was unreliable, so verbal affixes were excluded from the analyses.

[2] http://oracc.museum.upenn.edu/epsd2/sux; http://oracc.museum.upenn.edu/epsd2/index.html.

Japanese has a rich lexicon of two-character, two-morpheme Sino-Japanese compounds called *jukugo* (Ogawa and Saito, 2006; Joyce, 2013), as well as suffixing morphology on verbs. Two Python scripts were written that automatically parsed *jukugo* into their component morphemes and that separated verb roots from affixes. The Jukugo Database at `kanjidatabase.com` (Tamaoka et al., 2017) was used to help with *jukugo* parsing.

3.2 Phonographizing Sumerian and Japanese

The writing systems of both Japanese and Sumerian are considered highly logographic. However, while Japanese texts are typically written using a mix of logographic *kanji* and phonographic *kana*, any Japanese text can be written using *kana* alone. To ensure that we have a highly phonographic system for comparison, we included in our analyses a "phonographized" version of the Japanese data that treated all tokens as if they were written in *katakana*. For example, the morpheme *abura* "oil" is typically written as 油, but can be written in *kana* as あぶら or (rarely) アブラ; in our phonographized Japanese, *abura* is only written as アブラ.

For comparison, we also devised a phonographized version of Sumerian. Since Sumerian doesn't have a set of canonical phonographic characters in the way Japanese does, we constructed a hypothetical phonographized Sumerian using the following method: for each phonological form in the corpus, we found the spelling that most frequently mapped to that form. We then rewrote all instances of those phonological forms such that they were written with that most common spelling. This creates a system which would be considered minimally logographic under Sproat and Gutkin's distinct homophones treatment of logography. Since surface phonetic information isn't available for Sumerian, phonological forms were always treated as the cited dictionary transliterations, meaning that we treat each Sumerian morpheme as having only one possible phonological form.

Sumerian was further treated in two separate ways: the first way included all available (non-discarded) data, such that graphic form types included any character string which could map to a single morpheme. To get a sense of how logographic Sumerian is at a character level, we also ran the analyses on a subset of Sumerian that only included morphemes which could be written with a single character ("MonoChar Sumerian"). For example, MonoChar Sumerian would include morphemes such as  *i* "oil", but not morphemes such as  *šegin* "glue".

4 Results and Discussion

Results are given in Table 1. Even though the number of systems compared thus far is decidedly modest, the results are consistent with our expectations: for the more logographic systems, $I(G;M) > I(G;P)$, while the opposite is true for the phonographic systems.

The results are most salient for Japanese. The Sumerian results, while still in-line with expectations, are weaker in magnitude; whether or not this is a consequence of the smaller dataset or of an actual difference in how logographic written Sumerian was remains a point of future consideration. Additional points of comparison will be needed to get a more complete picture of whether the magnitude of the results (rather than just the valence) correlates with how logographic a system is.

With respect to the phonographized MonoChar Sumerian, the fact that $I(G;M) = I(G;P)$ makes sense given how phonographized Sumerian was crafted, and given that, for reasons stated earlier, each Sumerian morpheme only has one possible pronunciation form.

It is interesting to note that the scores for phonographized Sumerian are close to zero rather than strongly negative, i.e. written forms were about as informative about morpheme identity as they were about phonological form. This system could thus be viewed as more logographic than our phonographized Japanese, despite both systems being exemplar minimally logographic system in the distinct homophones sense. The ability to capture such a distinction marks another potential advantage of incorporating morpheme identity when quantifying logography.

5 Conclusion

The preliminary studies reported here offer a simple but promising means of measuring how logographic a writing system is. The inclusion of all three of phonological forms, graphic forms, and morphological identity paints a more complete picture of what it means for a writing system to be logographic. Ongoing work is focused on expanding these methods to a wider range of writing systems, as well as incorporating semantics.

Writing System	*I(G,M)*	*I(G,P)*	*l*	Writing System	*I(G,M)*	*I(G,P)*	*l*
Sumerian	6.950	6.819	**0.131**	Sumerian (**Ph**)	6.818	6.851	**-0.034**
Sumerian (**MC**)	6.228	6.074	**0.153**	Sumerian (**MC**, **Ph**)	6.111	6.111	**0.000**
Japanese	9.382	8.341	**1.041**	Japanese (**Ph**)	8.307	8.734	**-0.427**

Table 1: Results for the six different writing system variations. **MC** and **Ph** are abbreviations for "MonoChar" and "phonographized", respectively. See section 3.2 for details on the six variations.

Limitations

The work described here is part of an ongoing project, and our results, while promising, should be viewed as preliminary. We only report results for the writing systems of two languages, which is a major limitation for a study focusing on typology and cross-writing system variation; past studies in this vein (e.g., Marjou (2019); Sproat and Gutkin (2021); Rosati (2022)) have rightly considered a wider range of languages. While the systems we consider (including their "phonographized" versions) provide good points of comparison, the results would be strengthened by considering a wider range of writing systems (which the authors intend to do).

Finally, it should be noted that the morphological parsing done on the data used in this study may be imperfect, despite the first author's best efforts. Limitations in modern understanding of Sumerian result in some cases that should perhaps be viewed with some caution. Similarly, for Japanese, the treatment of all *jukugo* words as bimorphemic may or may not accurately reflect how such words should be analyzed in Modern Japanese. It's also possible that some non-*jukugo* two-*kanji* words were accidentally categorized and parsed as if they were *jukugo*. Certain non-*jukugo* compounds may have also escaped detection.

Given the in-progress nature of this research, code and (cleaned) datasets have not yet been made publicly available, but it is the authors' intention that these resources will be released in the future.

References

Susanne R Borgwaldt, Frauke M Hellwig, and Annette MB de Groot. 2004. Word-initial entropy in five languages: Letter to sound, and sound to letter. *Written Language & Literacy*, 7(2):165–184.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Peter T Daniels and William Bright. 1996. *The world's writing systems*. Oxford University Press on Demand.

Terry Joyce. 2013. The significance of the morphographic principle for the classification of writing systems. *Typology of writing systems*, pages 61–84.

Terry Joyce and Susanne R Borgwaldt. 2013. *Typology of writing systems: Introduction*, volume 51, pages 1–11. John Benjamins Publishing.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, 48:345–371.

Xavier Marjou. 2019. Oteann: Estimating the transparency of orthographies with an artificial neural network. *arXiv preprint arXiv:1912.13321*.

Piotr Michalowski. 2004. Sumerian. *The Cambridge Encyclopedia of the World's Ancient Languages*, pages 19–59.

Taeko Ogawa and Hirofumi Saito. 2006. Semantic activation in visual recognition of Japanese two-kanji compound words: Interference and facilitatory effects of neighbors. *Psychologia*, 49(3):162–177.

Gerald Penn and Travis Choma. 2006. Quantitative methods for classifying writing systems. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 117–120.

Athanassios Protopapas and Eleni L Vlahou. 2009. A comparative quantitative analysis of Greek orthographic transparency. *Behavior research methods*, 41:991–1008.

Henry Rogers. 2005. *Writing systems: A linguistic approach*, volume 18. Blackwell publishing.

Domenic Rosati. 2022. Learning to pronounce as measuring cross lingual joint orthography-phonology complexity. *arXiv preprint arXiv:2202.00794*.

Mark S Seidenberg and James L McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Masayoshi Shibatani. 1990. *The languages of Japan*. Cambridge University Press.

Noam Siegelman, Devin M Kearns, and Jay G Rueckl. 2020. Using information-theoretic measures to characterize the structure of the writing system: the case of orthographic-phonological regularities in English. *Behavior research methods*, 52:1292–1312.

Richard Sproat. 2000. *A computational theory of writing systems*. Cambridge University Press.

Richard Sproat and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528.

Katsuo Tamaoka, Shogo Makioka, Sander Sanders, and Rinus G Verdonschot. 2017. www.kanjidatabase.com: a new interactive online database for psychological and linguistic research on Japanese kanji and their compound words. *Psychological research*, 81:696–708.

Steve Tinney and Eleanor Robson. 2014. ORACC: The Open Richly Annotated Cuneiform Corpus.

Rebecca Treiman, John Mullennix, Ranka Bijeljac-Babic, and E Daylene Richmond-Welty. 1995. The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, 124(2):107.

Johannes C Ziegler, Arthur M Jacobs, and Gregory O Stone. 1996. Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods Instruments and Computers*, 28(4):504–515.

Johannes C Ziegler, Gregory O Stone, and Arthur M Jacobs. 1997. What is the pronunciation for -ough and the spelling for/u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods Instruments and Computers*, 29:600–618.