The Efficiency of Spelling-Sound and Spelling-Meaning Mappings in Two Logographic Writing Systems

Noah Hermalin

Abstract. A line of research in linguistics and cognitive science argues that many properties of language and linguistic structure are shaped, at least in part, by functional pressures to communicate efficiently, with efficiency defined relative to the trade-off between the accuracy and effort of communication. Research in this vein has argued that certain properties in language which seem intuitively inefficient, such as lexical ambiguity, are not only less inefficient than is commonly thought, but may in fact support, rather than contradict, the hypothesis that efficiency pressures shape language. The work presented here applies this line of thinking to two unrelated writing systems, written Sumerian and written Japanese, which both prominently feature properties which might be seen as inefficient, namely ambiguity and logography. Drawing on methods from information theory and past work on communicative efficiency, the efficiency of mappings between written forms and sound and between written forms and meanings is quantified for these two writing systems and compared against a battery of plausible hypothetical competitor systems, including variants which make greater use of phonography. The real systems are found to be more efficient overall than the hypothetical competitors, which is consistent with the efficiency hypothesis. The results are discussed relative to the question of efficiency pressures in writing systems, the inefficient 'reputation' of certain writing systems, and the linguistic fit of writing systems.

Introduction

Writing systems invariably have some degree of non-isomorphism between written forms and other linguistic information such as sounds and words. It is common, for example, for a writing system to have written forms that can be read in more than one way, or multiple ways of spelling a given sound or word. Another property found, to varying degrees, in virtually all writing systems is that a given word's spelling might not be a function only of that word's pronunciation; this is especially salient

Noah Hermalin 0009-0002-3200-5095
Department of Linguistics, University of California, Berkeley, US
E-mail: nmhermalin@berkeley.edu

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 12.
Fluxus Editions, Brest, 2026, pp. 551–595. https://doi.org/10.36824/2024-graf-herm
ISBN: 978-2-487055-12-4, e-ISBN: 978-2-487055-13-1

in writing systems that often have individual characters encode information at the lexical or morphological level. The former property, the lack of one-to-one mappings, can be termed *ambiguity*, *polyvalence*, or *non-isomorphism*. The latter property falls under the header of *logography*. Both of these properties have been cast by some as being sources of inefficiency in writing systems, and have factored into arguments both for and against the idea that writing systems are shaped by 'natural' pressures, such as pressures to be communicatively efficient. The research presented here is concerned with the interaction of these two properties, and how they affect writing system (in)efficiency. How, and why, do certain writing systems make use of polyvalent, logographic characters/spellings, and do these properties manifest in a way that facilitates, rather than impedes, communicative efficiency?

The expectation is that the presence and instantiation of logography and non-isomorphism in writing systems is more beneficial, rather than detrimental, to the efficiency of such writing systems, and thus constitute evidence in favor of the hypothesis that writing systems are shaped by communicative efficiency pressures. In other words, these properties, despite their uneven reputation, are not structural flaws which are merely tolerated for reasons of culture or a general 'conservatism' regarding writing systems. This hypothesis fits in not only with functionalist approaches to linguistics and cognitive science generally, particularly with regards to the burgeoning literature on how efficiency shapes language (Gibson et al., 2019), but also with broader question of why writing systems are they way they are (and aren't the ways they aren't).

Comments on Terminology

As is often remarked on, writing systems terminology varies by author. Appendix A provides brief definitions for certain terms as they will be used in the current paper.

Efficiency

Language, in all modalities, is used for communication. As such, one might expect, or at least hope, that language is well suited for communication. This expectation can be taken a step further by hypothesizing that language is the way it is (at least in part) *because* it has been shaped by communicative pressures. Adding one more layer of specificity, one can consider whether the pressure to communicate *efficiently*, as defined below, can explain various properties of human language. The hypothesis that efficiency shapes language will be referred to here as the *efficiency hypothesis*.

The last few decades have seen an influx of research that directly considers the efficiency hypothesis (for recent reviews, see Kemp, Xu, and Regier (2018), Gibson et al. (2019), Mahowald, Dautriche, et al. (2022), and Levshina (2022, pp. 3–35))[1]. Such research has focused on a range of linguistic topics, both broad and narrow, but much of it is unified by similar methodological approaches which make use of concepts and methods from information theory (Cover and Thomas, 1991; Shannon, 1948).

To begin, there is the basic information theoretic model of what communication is. As summarized by Gibson et al.:

> First, an information source selects a message to be transmitted. Next, the message is encoded into a signal, and that signal is sent to a receiver through some medium called a channel. The receiver then decodes the signal to recover the intended message. Successful communication takes place when the message recovered at the destination is [identical or nearly identical] to the message selected at the source. Efficiency in communication means that successful communication can be achieved with minimal effort on average by the sender and receiver. (Gibson et al., 2019, p. 391)

In the context of linguistic communication, both the information source and the encoder are a language user who wishes to linguistically convey some idea, message, knowledge state, etc. Similarly, the receiver and destination are both some other language user, who receives the signal and decodes it in an attempt to reconstruct the target message. The signal itself takes the form of linguistic units, such as words, as realized by a particular modality.

As mentioned above, communicative *efficiency* is based on both how accurate communication is and how effortful it is. Regarding the terminology that will be used in this paper, a system's *informativeness* is a matter of whether it enables communication which is accurate/successful. A system's *complexity* refers to how effortful it is to use that system. Ideally, a communication system is maximally informative and minimally complex: taken to the extreme, a maximally efficient system would be one that is completely effortless to learn and use and enables communication that is one-hundred percent accurate one-hundred percent of the time.

However, such a system does not exist. There is a tension underlying informativeness and complexity, namely that each of these two factors imposes limits on the other which leads to a trade-off relationship between them. In order for a communication system or event to increase

1. It should of course be acknowledged that functionalist, efficiency-based accounts of language are by no means the only accounts of language, nor is the efficiency hypothesis without good-faith critics. Some recent critiques of some efficiency-related arguments include Caplan, Kodner, and Yang (2020), Levshina (2021), and Morin and Koshevoy (2024).

informativeness, it typically needs to be more complex, and vice versa. In light of the push-pull of informativeness and complexity, efficiency becomes a matter of how well a system optimizes the informativeness-complexity trade-off. If a system is as informative as it can possibly be given its level of complexity and is as simple as it can be given how informative it is, it can be viewed as optimally efficient. An important consequence of this approach is that two systems that differ in both informativeness and complexity can still be comparably efficient, provided that the more complex system is more informative than the simpler system.

In order to put this model of efficiency into practice, it is necessary to define informativeness and complexity in ways that can be quantified. Conceptually, any measure which can compare the difference between the intended message of the sender and the reconstructed message of the receiver, or can capture the quality and consistency of using particular signals for communication, would be suitable for informativeness. Concepts from information theory (e.g., surprisal, entropy, mutual information) are often used to quantify informativeness, though there is not a single, unified informativeness metric which is the standard across the literature. Complexity (or effort) encompasses a wide range of factors relevant to language. Measures such as utterance length, rate of communication, motor difficulty, or short-term memory burdens are among the ways to capture aspects of in-the-moment effort. Counting the number of distinct unit types in a language (e.g., distinct morphemes, phonemes, etc.) or the number of grammatical rules that a language's users need to know would help measure effort from the perspective of acquiring and maintaining proficiency. Like with informativeness, there is no single, unified complexity metric, and any given element of complexity may itself be measured in different ways. In addition, some dimensions of complexity may trade-off with other dimensions of complexity.

The question of whether language is shaped by efficiency pressures can be rephrased as: to what extent are various *properties* of language shaped by efficiency pressures? Turning to the literature, one finds evidence that quite a few properties of language seem to be shaped at least in part by efficiency pressures, including (but not limited to): word length (e.g., Mahowald, Fedorenko, Piantadosi, and Gibson, 2013; Piantadosi, Tily, and Gibson, 2011; Pimentel, Meister, et al., 2023; Zipf, 1949); articulation (e.g., Aylett and Turk, 2004; Cohen Priva, 2017; Gahl, Yao, and Johnson, 2012); syntax and word order (e.g., Futrell, Mahowald, and Gibson, 2015; Gildea and Temperley, 2010); morphology (e.g., Dye, Milin, Futrell, and Ramscar (2018) for grammatical gender; Wang and Walther (2023) for classifiers; Fedzechkina, Jaeger, and Newport (2011) and Lee (2009), and Kurumada and Jaeger (2015) for case marking; Mollica, Bacon, Zaslavsky, et al. (2021) for tense); pronouns (Denić, Steinert-Threlkeld, and Szymanik, 2020); information rate (e.g., Coupé,

Oh, Dediu, and Pellegrino, 2019; Jaeger, 2010); and, as will be discussed below, semantic typology, colexification, and ambiguity.

Ambiguity and Efficiency

One prevalent property of language which is often raised as an example of communicative *inefficiency*, is that of ambiguity, which is present at multiple levels of linguistic structure (cf. Wasow, 2015). One of the more salient types of ambiguity is lexical ambiguity, including homophony and polysemy/colexification, and it is lexical ambiguity that will be the main focus of this brief review.

Intuitively, ambiguity is detrimental for communication from the perspective of the decoder who, in the absence of additional information, will have a reduced likelihood of reconstructing the intended message; in terms of a complexity-informativeness trade-off, ambiguity reduces informativeness, and as such reducing ambiguity has the potential to improve efficiency. Indeed, there is evidence that ambiguity avoidance may be an explanatory factor in language change (e.g., Wedel, Kaplan, and Jackson (2013) regarding phonological mergers; Zehentner (2022) on the emergence of syntactic constructions) and acquisition (e.g., Yin and White, 2018). There is also evidence that, at least for high frequency forms, homophones are less common than would be expected by chance (Trott and Bergen, 2020; 2022), while minimal pairs are more common than would be expected (see also Dautriche et al., 2017).

Nonetheless, it has been proposed that some amount of ambiguity is necessary (e.g., Juba, Kalai, Khanna, and Sudan, 2011) and may even be beneficial for language (e.g., Peloquin, Goodman, and Frank, 2020; Piantadosi, Tily, and Gibson, 2012). The necessity of ambiguity is intuitive when one considers how impractical it would be to have a completely unambiguous system where every possible meaning has its own distinct form. The crux of the pro-efficiency argument for ambiguity is that ambiguity has the potential to decrease system complexity, by virtue of decreasing the number of distinct forms needed to communicate, and that in practice, contextual information will serve to disambiguate otherwise ambiguous forms, mitigating losses to informativeness. Note that this line of argument isn't incompatible with the view that shifts away from ambiguity can also contribute to efficiency. Ambiguity isn't always good or always bad, but rather is a property which languages can take advantage of to increase efficiency, but will shy away from if the cost to informativeness is not worth the decrease in complexity. For example, building on their previous finding that phonological mergers are avoided if they would lead to greater homophony in the lexicon, Wedel, Jackson, and Kaplan (2013) temper their earlier results by showing that additional contextual information can influence merger avoidance; for

example, mergers that would result in homophones that share a lexical category (e.g., both nouns) are less common than mergers that would result in cross-category homophones (e.g., a noun and a verb).

The potential efficiency of lexical ambiguity has been approached from various angles. Corpus studies have found that lower effort forms are more likely to have a disproportionately large number of meanings (Piantadosi, Tily, and Gibson, 2012), and a word's predictability in context has a positive correlation with how lexically ambiguous it is (Pimentel, Maudslay, Blasi, and Cotterell, 2020), which suggests that language users may make contexts more informative when using words which are more lexically ambiguous. Evidence from communication simulations also support the efficiency argument, with Peloquin, Goodman, and Frank (2020) finding that the presence of ambiguous forms were more likely in situations with greater context.

Turning to the semantics of ambiguity, there are the questions of whether certain meanings are more likely to colexify, (i.e., share a form), and whether typological trends in how meanings colexify are consistent with the efficiency hypothesis. Studies on semantic typology have yielded evidence that the lexicons of natural languages do indeed seem to 'carve up' the space of possible meanings in ways that facilitate efficient communication (Kemp, Xu, and Regier, 2018). Such results have been found for a range of semantic domains, including color (e.g., Zaslavsky, Kemp, Regier, and Tishby, 2018), kinship (Kemp and Regier, 2012), containers (Xu, Regier, and Malt, 2016), animals (Zaslavsky, Regier, Tishby, and Kemp, 2019), sensory terms (Majid et al., 2018; Winter, Perlman, and Majid, 2018), and terms relating to climate and environment (Regier, Carstensen, and Kemp, 2016). A recurring notion in many of these studies is the idea that languages adapt to be efficient relative to the cultural and environmental context in which they are used. In particular, the degree to which communicating a particular concept will affect the overall efficiency of the language will be based in part on the likelihood with which users of that language actually need to communicate about that concept. Even if communicating about a particular concept is costly, it will have a minimal effect on efficiency overall so long as communicating about that concept is a rare occurrence. If the distinction between two concepts rarely matters, those concepts can more safely be colexified.

Expanding the scope beyond individual semantic domains, Xu, Duong, et al. (2020) consider which general factors influence colexification rates. Their results, based on data from 246 different languages, suggest that both conceptual similarity (e.g., *fire* is similar to *flame* are similar concepts) and conceptual relatedness/associativity (e.g., *scent* is associated with *nose*) are strong predictors of whether two meanings will colexify. Similar conclusions were made by Karjus et al. (2021), who found that semantic similarity predicted colexification in a novel language

learning participant study, but also found that participants were less likely to colexify similar concepts that often needed to be distinguished between during communication. Brochhagen and Boleda (2022) expand on the role of similariy in colexification, arguing that, while meanings are more likely to colexify if they're semantically related, languages may avoid colexifying meanings that are *too* similar. They found that pairs of extremely related concepts were less likely to colexify than pairs of more moderately related concepts, and proposed that, as meanings become too closely related, the risk of confusability increases, and thus colexifying such meanings risks hurting informativeness to an extent that is not worth the modest reduction in complexity.

To summarize, there is mounting evidence that languages, and language users, make use of ambiguity in ways that reduce complexity at little or no cost to informativeness, thereby supporting efficient communication. One might expect that, excluding spoken language-specific topics such as phonology, much of the previous work on lexical ambiguity and efficiency directly translates to written language. However, this rests on the assumption that written language is always a one-to-one mirror of spoken language, which is not the case. Sharing a spoken form is no guarantee of sharing a written form (e.g., English ⟨tea⟩ versus ⟨tee⟩, both /ti/), and vice versa (e.g., English /wɪnd/ and /wa͡ɪnd/, both ⟨wind⟩). Even different senses of a polysemous word may have distinct forms depending on the modality: for instance, (spoken) Japanese /atu/ "hot" is spelled 熱 in reference to hot objects, but as 暑 for hot weather or climate. As such, whether or not writing systems and written language make efficient use of ambiguity is a question that can't be answered without focusing directly on writing systems and written language[2].

Writing Systems and Efficiency

The idea that writing systems are (or are not) shaped by functional communicative pressures is not new. Many arguments regarding writing system efficiency could be slotted into three loose categories: either functional pressures exert little to no influence over writing systems;

2. It is true that some the past work on the efficiency of lexical ambiguity has used written data (such as text corpora or language simulations that use written communication, as in Karjus et al. (2021). In most cases this is done for reasons of data availability/convenience, and the use of written data may be couched as being a proxy of spoken language (and, when possible, written data may be converted to approximate spoken data). Nonetheless, even though such studies typically aren't framed as being about written language per se, they are still relevant to whether written language is efficient (and in fairness, they sometimes are treated as being analyses of written rather than spoken language by virtue of their choice of data, though this may get portrayed as a concession rather than a goal, as in e.g., Pimentel, Nikkarinen, et al., 2021).

or functional pressures act on writing system in such a way that some (types of) writing systems end up more efficient than others; or functional pressures act on writing system in such a way that all writing systems are comparably efficient. What follows is a brief review, with an eye for the main topics of the current study, namely ambiguity and logography.

The Validity of Functionalist Approaches to Writing Systems

To begin, it is worth addressing the possibility that writing systems are not shaped by functional pressures. In short, this view often takes the form of arguing that writing systems are the way they are mostly, if not entirely, because of societal factors (e.g., (Coulmas, 2003; Trigger, 1998; Cooper, 2004, p. 91; Coulmas, 2009, p. 9; de Voogt, 2024, p. 896; see also discussion in Houston, 2012a), and such societal pressures often serve to reduce or inhibit shifts towards writing system utility (e.g., Hannas, 1997). A succinct form of this argument is provided by Coulmas:

> It is a mistake to see writing systems as quasi-natural organisms governed in their development by natural laws. Every script is a cultural implement subject to human ingenuity and error, created under certain circumstances for certain purposes and a certain language. To be sure, there are common traits, and economy of effort clearly is one of the guiding principles of human behaviour. Yet there is plenty of room for waste, extravagance and manifestations of the human mind defying bare utility. Cultural inertia and conservatism [...] and normativism [...] are strong forces at work in every literate community. They have little to do with writing systems as such or with their efficiency, yet they exercise a strong influence on their formation. Writing is a cultural product *par excellence*, and its development must be understood as such rather than in quasi-naturalistic terms. (Coulmas, 2003, pp. 200–201)

Written language does indeed seem to be more resistant to diachronic shifts than non-written language, especially if that written language is standardized, or functions as a prestige literary language. Top-down spelling reforms are, of course, a result of social, cultural, and political action, and a lack of spelling reforms (for writing systems that some might feel are in need of some changes) may be attributed to sociocultural factors. To argue that writing systems are completely unaffected by functional pressures, however, would be premature[3]. The question is: *to*

3. Two points must be made here: first, it is rare to encounter an argument as extreme as "writing systems are completely unaffected by functional pressures"; as evidenced in the above quote from Coulmas, the potential for functional pressures to shape writing systems can be acknowledged even when arguing for the relative strength of social and cultural pressures. Second, arguments that cast doubt on the role of functional pressures are often made in response to proposals that writing systems are *predominantly* shaped by such pressures (such as the 'evolutionary' views of writing

what extent do functional pressures shape writing systems? Even if 'natural' pressures only had a weak effect on writing systems, such pressures would still very much worth investigating, and any results in such a vein would absolutely be relevant to broader questions of human language, writing systems, and cognition. Or, as neatly summarized by Sampson:

> It is understandable that political considerations may sometimes outweigh simple efficiency in deciding what form of writing a society uses. One might suppose, though, that where politics does not impinge, functional considerations would constrain the ways in which a script could evolve. (Sampson, 2015, p. 58)

Criteria of Writing System Effectiveness

There have been a range of proposals over the years regarding what criteria best measure the effectiveness of writing systems (e.g., Baroni (2011), Cahill and Karan (2008), Coulmas (2009), Daniels and Share (2018), Powlison (1968), Rogers (1995), and Venezky (1970); see also Meletis (2019, p. 206)). Such proposals vary as to whether or not these measures of effectiveness *predict* how writing systems develop, or are an ideal (attainable or otherwise) to which writing systems should aspire. Considerations such as regularity of mappings, inventory size, and visual distinctiveness, which are among the more frequently proposed criteria, could be treated as measures of either informativeness or complexity, as per the current study's treatment of efficiency. Notably, such lists also tend to make mention of ambiguity, specifically as something to be avoided by writing systems, though it is sometimes acknowledged that writing systems may need to balance one source of ambiguity with another (e.g., morphological regularity versus phonetic regularity). Indeed, the theme of trade-offs, in line with the informativeness-complexity trade-off that is at the core of how the current study treats efficiency, is not uncommon in the writing systems literature (e.g., Civil, 1973; Daniels and Share, 2018; Rogers, 1995; Salomon, 2012; Seidenberg, 2012; Share, 2014; Trigger, 1998).

Inventory size and learnability are two factors which have featured prominently in some of the more influential (and critiqued) theories regarding writing system development. While not the first to raise such arguments, Gelb (1963), and contemporaries such as Diringer (1962), helped popularize the idea that writing systems necessarily 'evolve' towards greater efficiency, following a monotonic path from logographic

---

system development reviewed in this section), and in fairness it is indeed unlikely that functional pressures are the only force behind writing system development and structure.

to syllabic to segmental/alphabetic. By Gelb's account, a writing system's efficiency is based on his 'principle of economy,' defined as "the effective expression of the language by means of the smallest possible number of signs." (Gelb, 1963, p. 72). In other words, efficient writing systems are those that prioritize reducing inventory size (complexity), not those that optimize the informativeness-complexity trade-off (cf. the current study's treatement of efficiency).

Such arguments regarding writing system evolution and the comparative utility of different writing system types (in part or in whole) have faced no shortage of critics (e.g., Baroni, 2011; Coulmas, 2003; Daniels, 1990; de Voogt, 2024; Mattingly, 1985), but the general idea that some (types of) writing systems are more functional or practical is still present, in various forms, in research on writing systems.

The Effectiveness of Logography

Among proponents of the view that some types of writing systems are inherently (dis)advantageous, the most common stance (at least in the Anglophone literature) has been that logography is less efficient than phonography. Points raised in support of more logographic writing systems being less practical than more phonographic systems include: logographic systems require larger character inventories (e.g., Gelb, 1963; Pulgram, 1976); logographic systems are inconvenient for lexicography (DeFrancis, 1996); logographic systems don't capitalize on learner's already strong phonological knowledge (DeFrancis, 1996; Hannas, 1997); logographic systems take more time and effort to acquire and master (Hannas, 1997; Powlison, 1968); and logographic characters tend to be visually complex. Such discussions are usually focused on written Chinese (or Sinograms generally, including as used in written Japanese and Korean), though similar ideas can be found (or are at least mentioned) in reference to the writing systems of ancient Egypt (e.g., Baines (2004, p. 181)) and Mesopotamia (e.g., Cooper (2004, p. 92); N. Veldhuis (2011, p. 86)), or early writing systems more generally (Woods, 2010, p. 22).

There is, of course, not unanimous agreement that logography is disadvantageous, and indeed the current study's hypotheses are at odds with the idea that logography is inherently inefficient. This is not because of a lack of evidence for many of the proposed downsides of logographic systems: in more logographic systems, characters do tend to be more visually complex (Chang, Chen, and Perfetti, 2018; Miton and Morin, 2021); character inventories are larger (e.g., Chang, Chen, and Perfetti, 2018); characters are less informative about phonological information (e.g., Marjou, 2019; Sproat and Gutkin, 2021); and total literacy (at least regarding character knowledge and formal schooling) takes longer to acquire. A critical point to make is that none of these points necessarily lead to inefficiency, as defined here. Rather, they suggest

that logographic systems will trend towards increased complexity (regarding visual complexity and memory/acquisition) and decreased informativeness regarding sounds. However, such systems are only inefficient if these downsides aren't accompanied by improvements in other areas. Indeed, certain advantages have been proposed for retaining logography, such as: accommodation of dialectal differences (e.g., Sampson (1985, p. 170); Chen (1999, pp. 140–141)); homophone disambiguation (e.g., Sampson (1985, p. 170)); Taylor and Taylor (1995, p. 323); Chen (1999, pp. 140–141); Handel (2013); Meletis (2019, p. 268)) regularity of spellings with regards to morphological alternations (e.g., Berg and Aronoff, 2017; Chomsky and Halle, 1968; Venezky, 1970); semantic transparency (e.g., Handel, 2013); and ease of scanning texts for specific words or concepts (Cooper, 2004, p. 98); cf. Morita and Tamaoka, 2002).

Linguistic Fit

Another possibility is that certain types of writing systems are better for certain types of spoken languages (e.g., Baroni, 2011; Frost, 2012; Halliday, 2010; Handel, 2019; Joyce and Meletis, 2021; Katz and Frost, 1992; Mattingly, 1985; Meletis, 2019; Seidenberg, 2011; 2012; Voegelin and Voegelin, 1961), a concept known (among other names) as linguistic fit. Most takes on the linguistic fit argument can be divided into two pieces, one analytical and one empirical: the analytical is that certain types of writing system are better suited to certain types of spoken languages, particularly with regards to that spoken language's morphology and phonology (e.g., simpler syllable structure is more compatible with syllabic writing systems). The empirical portion of the linguistic fit argument is whether it is indeed the case that writing systems trend towards, or away from, certain structural properties so as to better fit their respective spoken languages. Proponents of linguistic fit will mostly agree that the empirical question can be answered in the affirmative (though see Meletis (2019, pp. 355–356) for a more nuanced answer), but there is disagreement over whether this fit is *optimized* (see (responses to) Frost, 2012).

Ambiguity

Ambiguity avoidance has factored in to various discussions of writing system development (e.g., Boltz, 2011; Woods, 2010), borrowing (e.g., Handel, 2019), and reform (e.g., Chen, 1999; Handel, 2013). Innovations and developments such as determinatives, using different spellings for homophonous words, and including redundant information in spellings have been touted as evidence that writing systems prefer to avoid ambiguity. However, the inevitability and trade-offs of ambiguity in writing systems have been noted as well (e.g., Fortson, 2023; Salomon, 2012),

and the presence of ambiguity in writing system has also been cited as evidence against such systems being shaped by functional pressures (e.g., de Voogt, 2024). What is missing from many of these claims, however, is quantified empirical evidence: for example, claims about the disambiguating benefits of determinatives tend not to be accompanied by counts of just how many words are disambiguated by the use of determinatives.

Quantitative methods are more common in the robust literature on the ambiguity and consistency of spelling-sound mappings in writing systems, often referred to as *orthographic depth* or *transparency* (e.g., Borleffs, Maassen, Lyytinen, and Zwarts, 2017). While not typically focused on efficiency or functional pressures, the depth/transparency literature overlaps with the current study not only in its focus on ambiguity, but also its use of information theoretic methods (as in e.g., Borgwaldt, Hellwig, and Groot (2004) and Protopapas and Vlahou (2009); and Siegelman, Kearns, and Rueckl, 2020). Such studies have predominantly focused on more phonographic (usually more alphabetic) systems, however, and the emphasis on character-sound mappings has left underexplored the topic of character-morpheme mappings (though see Siegelman, Rueckl, et al., 2022).

Cognitive and Quantitative Approaches to Writing System Efficiency

Finally, is worth highlighting some recent studies that, like the current study, consider writing systems from a more cognitive, quantitative, and/or efficiency-focused (in the *efficiency hypothesis* sense) perspective. Much of the research in this vein has focused on visual properties, such as visual complexity and redundancy (Changizi and Shimojo, 2005; Han, Kelly, Winters, and Kemp, 2022; Kelly, Winters, Miton, and Morin, 2021; Koshevoy, Miton, and Morin, 2023; Miton and Morin, 2021), representational iconicity (e.g., Kuperman and Bertram (2013) for the spelling of compound words), and congruence with human visual cognition (e.g., Changizi, Zhang, Ye, and Shimojo, 2006; Morin, 2018). Overall, the results of these studies support the idea that functional pressures shape the development of written character forms. Regarding visual simplicity, the available evidence suggests that new written forms simplify rapidly over a few generations (Garrod et al., 2007; Kelly, Winters, Miton, and Morin, 2021), but over longer stretches of time characters don't seem to trend towards greater simplicity (Miton and Morin, 2021), and may even trend towards increased visual complexity, perhaps to support visual distinctiveness (Han, Kelly, Winters, and Kemp, 2022).

Case Studies—Sumerian and Japanese

Both ambiguity and logography have somewhat uneven reputations regarding their contributions towards the utility of writing systems and written language. Logography has traditionally been considered a negative, and while this view is increasingly called into doubt, there remains a demand for more quantitative, empirical investigations of logography's advantages and disadvantages. Such approaches have been used to test the proposed efficiency of ambiguity, but rarely with a focus on writing systems and written language. The current study aims to help fill both of these gaps.

In light of this focus on ambiguity and logography, it makes sense for the initial case studies to be two writing systems which a.) make extensive use of logography (including more morphographic and semantographic spellings) and b.) feature numerous polyvalent characters and spellings: written Sumerian and written Japanese. If it's found that even these highly ambiguous, highly logographic systems are communicatively efficient, it would be strong evidence that these properties are not as inefficient as has previously been assumed, which in turn would support the general hypothesis that writing systems are shaped by efficiency pressures.

What follows is a quick overview of the relevant similarities and differences between (written and spoken) Sumerian (as it was used in the Ur III period, c. 2112-2004 BCE) and Japanese.

Morphology and Phonology

Spoken Sumerian and spoken Japanese can both be generally described as agglutinating: morpheme boundaries are usually fairly transparent, compounds can be formed by concatenating two (or more) roots, and most grammatical information is represented via affixes or clitics.

Syllables in both languages are typically analyzed as minimally V, maximally CVC (depending on the analysis, the Japanese syllable may be maximally CyVC/CyVV or CyVVC)[4]. Sumerian is less restrictive than Japanese regarding which consonants can occur as syllable codas. In addition to having similar syllable structures, the two languages are comparable in the size of their respective phoneme inventories[5].

4. The palatal glide is rendered in this paper as /y/ rather than IPA /j/.

5. Sumerian phoneme inventory: /p, b, t, d, k, g, s, z, š, h, m, n, g̃, l, r/; /a, i, e, u/. Among the possible Sumerian phonemes which aren't included here are the glottal stop and controversial /r̂/ phoneme (see also Black, 1990). They are omitted because of their absence (or inconsistent usage) in the text corpus used as data for the current study. It is unclear whether Sumerian had phonemic vowel length (compare e.g., Jagersma (2010) with Michalowski, 2020), but the current study will assume that all

In Sumerian, most non-borrowed morphemes are monosyllabic, and some grammatical morphemes may be analyzed as being shorter than a syllable. Most borrowed morphemes (of Semitic origin) are multisyllabic. Japanese morphemes of Japanese origin (*wago*) tend to be one or two syllables (and one or two morae) in length with (C)V or (C)VCV forms, though many verb stems are underlyingly CVC. Japanese morphemes of Chinese origin (*Sino-Japanese*), which constitute a large percentage of the lexicon, tend to be monosyllabic but bimoraic. More recent (mostly European-origin) borrowings into Japanese (*gairaigo*) span a range of different lengths.

The two example sentences given below illustrate some of the morphological similarities (and differences) between Sumerian 1 and Japanese 2. For clarity, content morphemes are formatted in bold.

(1) **dub**-**sar**-e **e**-**gal** **dumu**-ani-ra
tablet-**write**-ERG **house**-**great** **child**-3S.A.POSS-DAT
mu-na-n-**du**
VENT-3S.A.DAT-3S.ERG-**build**

"The scribe built a palace for their child."

(2) **gaku**-**sya**-ga **siro**-o **oya**-ni **tate**-te-**age**-ta
academic-**person**-NOM **castle**-ACC **parent**-DAT **build**-**give**-PAST

"The scholar built a castle for their parent(s)."

Writing Systems

To begin, the two writing systems make use of both logographic and phonographic mappings at the character level: a character can map to an entire morpheme (or, more rarely, multimorphemic word), or can be used to encode a stable portion (no smaller than a syllable/mora) of a phonological form. In Sumerian, any given character may be used both logographically and phonographically: for example, 𒀀 can be used logographically to write either *a* "water" or *duru*$_5$[6] "wet," but 𒀀 can also

Sumerian vowels are short, again for reasons relating to the available data.

Japanese phoneme inventory: /p, b, t, d, k, g, s, z, h, m, n, ɾ, w, y, N/; /a, i, u, e, o/. Two caveats to note: First, vowel length is contrastive in Japanese, so it may be more accurate to list the vowel phoneme inventory as /a, i, u, e, o, aː, iː, uː, eː, oː/. Second, each consonant phoneme (other than /w, y, N/) has a contrastive palatalized counterpart, though such palatal consonants could be analyzed as Cy sequences rather than individual C^{y} consonants.

6. Some brief notes on formatting and transliteration practices for Sumerian: Each character has a standardized name, which is always formatted in all caps. To avoid two distinct characters having the same name, a character that would have the same name

TABLE 1. Examples of polyvalent characters in the two target writing systems.

Sumerian			Japanese		
Character	Sound	Meaning	Character	Sound	Meaning
𒀀	a	water	食	ta(be)	to eat
𒀀	duru	wet	食	syoku	meal, eating
𒄑	ğidru	scepter	食	ku(w)	to eat
𒄑	pa	stick, branch	生	u(m)	birth/born
𒄑	sag	to strike	生	nama	raw, fresh
𒄑	ugula	overseer	生	sei	life

be used to encode the syllable /a/, showing up in spellings such as 𒀀 *-a* "-LOC," 𒀀𒅖 *-a-ni* "-3S.A.POSS," and 𒀀𒂍 *a-ga* "rear, back". Written Sumerian does not have a set of characters which are *only* used as phonograms: for any given character, that character will likely have at least one logographic function.

Japanese, by contrast, has two sets of characters which are used exclusively as moraic phonograms: *hiragana* and *katakana*, collectively referred to as *kana*. The two types of kana have different usage patterns (though kana usage as a whole is somewhat flexible). For each hiragana character, there is one katakana character with the exact same phonographic value, and vice versa. The other set of characters traditionally used in Japanese are *kanji*, which are typically used as logograms. Modern written Japanese also makes use of the Roman alphabet *rōmaji*.

Sumerian and Japanese spellings and character can be ambiguous from both the reader's and writer's perspective. In both writing systems there are numerous characters which can each map to multiple different morphemes; these morphemes are often, but not always, semantically related. See table 1 for examples. In addition, as already mentioned, a character in Sumerian can have a mix of different logographic and phonographic mappings. Regarding encoding ambiguity, in both writing systems any given sound string and/or morpheme may have multiple valid spellings. For example, Japanese *oka* "hill' can be spelled logographically as 丘 or 岡, and phonographically as おか (using hiragana) or オカ (using katakana). When spelled logographically, spellings are usually unaffected by surface phonological variation, but phonographic spellings may vary depending on surface forms.

---

as another character has a subscript number appended to its name (so, for example, KA and KA_2 refer to different characters). Unless the reading of a character is unknown, Sumerian words are typically transliterated at the level of the characters' readings (i.e., approximate pronunciations). In this transliteration, hyphens indicate character boundaries, not morpheme boundaries, and subscript numbers serve as markers which indicate character identity.

One last striking similarity present in the two writing systems relates to the spellings of bound grammatical morphemes. Vowel-initial affixes are present in both spoken Sumerian and spoken Japanese. When these affixes immediately follow a consonant-final morpheme, they may be spelled as if they begin with the preceding morpheme's final consonant. For example, the Sumerian nominalizing suffix /-a/[7] has a default spelling of 𒀀 *a*, but is often written as 𒋛 *sa* after /s/ (e.g., 𒍑𒋛 *us$_2$-sa* "follow-NOMINALIZER"), 𒊭 *ra* after /r/ (e.g., 𒀜𒊭 *sar-ra* "write-NOMINALIZER"), and so forth. While this spelling strategy is used for both nominal and verbal morphology in Sumerian (albeit inconsistently), it is more or less exclusive to, and mandatory for, verbal suffixes in Japanese. For example, the non-past suffix, which is underlyingly /-u/ (after consonant-final stems), is written as む *mu* in 飲む /nom-u/ "drink," as す *su* in 話す /hanas-u/ "speak," and so forth.

This spelling strategy has two immediate consequences for ambiguity in the writing system. First, it exacerbates non-isomorphism between characters and grammatical morphemes. For Sumerian, which doesn't have dedicated phonograms in the way Japanese does, this can also serve to inflate the number of morpheme mappings of individual characters. However, the overt, redundant indication of the preceding morpheme's final consonant may serve to disambiguate what would otherwise be an ambiguous written form: for example, Sumerian 𒄑 is, in a vacuum, highly ambiguous, with values including (but far from limited to) *gub* "to stand," *tum$_2$* "to carry," *g̃en* "to go," and *kur$_x$* "to enter". When suffixed, such as with the aforementioned nominalizing /-a/, this ambiguity can be abated by spelling the suffix as if it were /-Ca/ rather than /-a/: using 𒁍 *ba* would indicate that the target is b-final (*gub*), using 𒈬 *ma* would indicate that the target is n-final (*g̃en*), etc. The disambiguating potential of this way of spelling affixes has already been noted elsewhere (e.g., Sampson (1985, p. 85); Coulmas (2003, p. 182); Handel, 2019; Honda, 2007; 2008), though the results of Honda (2007) suggest that, at least for Japanese, the overall reduction in ambiguity might be more modest than expected.

In summary, the two writing systems, and their respective spoken languages, have substantial structural overlap and make substantial use of both logography and ambiguity (see also Handel (2019, pp. 281–308) for similar observations on the potential for comparison between Sumerian and Japanese). These similarities can't be attributed to any genealogical or sociocultural relation, since Japanese and Sumerian (written and spoken) are completely unrelated, with their usage separated by thousands of kilometers and thousands of years. In addition, the two writing systems, as analyzed here, have had a comparable amount of 'development' time: Ur III Sumerian is roughly 1,000-1,400 years removed

7. Perhaps underlyingly /-ʔa/ (Jagersma, 2010).

from the earliest extant cuneiform texts (Woods, 2006), while modern written Japanese is roughly 1,200 years removed from the earliest clear evidence of vernacular written Japanese (Frellesvig, 2010; Seeley, 2000). As such, any observed similarities between the two system can't be attributed to their simply being related, and any observed differences between the two systems can't be attributed to one system having had less time to be shaped by efficiency pressures.

Data

The Sumerian data were drawn from the Ur III Administrative Corpus (hereafter Ur3Admin), hosted by the electronic Pennsylvania Sumerian Dictionary 2 (ePSD2), a project of the Open Richly Annotated Sumerian Corpus (ORACC; Tinney and Robson, 2014)[8]. As suggested by the name, the Ur3Admin corpus consists solely of administrative documents dated to the Ur III period (c.2112-2004 BCE); these documents are mostly receipts and records of debts, wages, property, taxation, resource allocation, and trade. The Ur3Admin corpus was chosen because it is the largest single-genre, single-time period annotated corpus available for written Sumerian.

The Japanese data were drawn from the Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa et al., 2014). BCCWJ draws on written material published between 1976 and 2006, and includes texts from a variety of sources and genres, such as newspapers, books and textbooks, magazines, online message boards, blogs, and official government minutes.

Both BCCWJ and ORACC are organized primarily at the word level. Because the current study is concerned with morphemes, rather than words, additional processing needed to be done to get morpheme-level information and mappings. In particular, compound words needed to be parsed into their constituent morphemes, and affixes needed to be separated from roots, in such a way that individual morphemes can be aligned with (minimally character-length) portions of written forms. Morphological and character-level parsing and aligning were mostly handled by custom Python scripts written for the task, with additional manual adjustments as needed.

During cleaning, certain word tokens and types needed to be removed. For the Sumerian data, proper nouns, most numerals, words with uncertain pronunciation or meaning, and tokens marked as damaged were all removed. For the Japanese data, recent foreign borrowings,

8. At time of writing, raw downloads for the Ur3Admin corpus can be found here: `https://oracc.museum.upenn.edu/epsd2/JSON/index.html`

TABLE 2. Type and token counts after additional cleaning and processing

	Character Tokens	Morpheme Tokens	Spelling Types	Morpheme Types
Japanese	55.49M	45.65M	4826	2531
Sumerian	2.83M	2.36M	2281	1716

proper nouns, and lower frequency words and forms were all excluded[9]. Even after cleaning, the size of the Japanese data was many times larger than the Sumerian data (see table 2).

The data was organized relative to spelling-sound-morpheme triplets. The set of mappings between spellings and sounds/meanings was determined based solely on these triplets, and was independent of whether that spelling would traditionally be considered to be logographic or phonographic. For example, based on the data snippet given in table 3, the spelling DU (𒁲) has (at least) three spelling-sound mappings (ra, gub, de) and (at least) four spelling-morpheme mappings (/-a/ "NOMINALIZER," /-ak/ "GENITIVE," /gub/ "to stand," /de/ "to carry").

TABLE 3. Snippet of the organized, cleaned Sumerian data

Graphic form	Pronunciation	Morpheme/meaning	Frequency
SI-IGI	silim	/silim/ "healthy"	2
DI	silim	/silim/ "healthy"	13
DU	ra	/-a/ "NOMINALIZER"	902
DU	ra	/-ak/ "GENITIVE"	8
DU	gub	/gub/ "to stand"	7464
DU	de	/de/ "to carry"	1060

9. The decision to remove recent foreign borrowings (*gairaigo*) from the Japanese data was made for a few reasons. First, *gairaigo* words are almost always written phonographically, using katakana, with logographic spellings being exceedingly rare. In addition, the fact that *gairaigo* are *recent* borrowings (mostly from the past century or so) means that they were absent from most of the history of written Japanese, and it's possible that the efficiency of the writing system hasn't 'caught up' to the influx of recent borrowings (though note that past findings such as those of Kelly, Winters, Miton, and Morin (2021) suggest that efficiency pressures could operate on a writing system, or at least on written forms, in time spans as short as a couple centuries). More importantly, the current study is most directly concerned with questions of non-isomorphism and logography, meaning that *gairaigo*, which are almost always written phonographically, are of less relevance. It should nonetheless be stressed that removing *gairaigo* does mean that the Japanese data used here should *not* be treated as representative of modern written Japanese, which makes extensive use of *gairaigo*, but rather as a proxy for written Japanese before the recent influx of borrowing from (mostly) English and other European languages.

Methods

Complexity and Informativeness

For this study, two simple complexity measures were used to quantify system effort: *strokes per morpheme* and *characters per morpheme*. The former measure serves as a direct approximation of motor production effort, while also providing a measure that at least indirectly indicates visual complexity (Chang, Chen, and Perfetti, 2018). Because different characters tend to take up a comparable amount of space within a system, the character-based metric captures the spatial and material cost of a text. For Sumerian, character stroke counts were manually defined for each character based on the exemplar forms given by Steve Tinney's Cuneiform Composite font[10]. For Japanese, each character has a canonical stroke count.

Following past work on efficiency pressures in language, (especially Kemp and Regier (2012) and Xu, Regier, and Malt (2016), and Kemp, Xu, and Regier, 2018) the informativeness metric used here is based on the expected amount of information that is lost during communication, here termed *communicative cost* (ComCost, or C). Recall that an information theoretic model of communication involves a sender (writer) encoding messages into signals which are decoded by a receiver (reader), who attempts to reconstruct the original message. These messages can be modeled as probability distributions over possible values. In the current study, these values are either phonological forms or morphemes, and each system will thus have two different informativeness scores: the phonological communicative cost captures how informative a system is about sounds, while the semantic communicative cost captures how informative a system is about meanings/morphemes. Each of the informativeness measures will be compared against each of the two complexity measures, resulting in four different efficiency analyses for each system.

The instantiation of communicative cost used here makes the assumption that the writer knows exactly which value they are trying to communicate: effectively, the sender's original message is a probability distribution with all of the mass allocated to the target value. The reader's probability distribution over values given a written form is taken here to be based on how similar possible values are to the values which are empirically spelled using that written form. The assumption is that the reader expects similar values to have similar forms. For example, if a written form is often used to encode the syllable /ta/, the reader's probability distribution for that form will assign much of its probability mass to /ta/, but will also assign some mass to similar syllables (e.g., /da/, /ra/), while assigning very little mass to a highly dis-

10. See https://oracc.museum.upenn.edu/doc/help/visitingoracc/fonts/.

similar syllable like /bi/. Similarly, if a written form is often used to encode a morpheme which means "young," then the reader distribution for that form will assign more probability mass to semantically similar or related morphemes (e.g., "junior" or "child") than to semantically dissimilar morphemes (e.g., "potash").

Communicative cost, then, can be measured based on the difference between the writer's original probability distribution and the reader's reconstructed probability distribution. Again in line was past work, and in light of the assumption of writer certainty, this can be measured as the reader's surprisal during communication about a given value. However, while it's is assumed that the reader incorporates semantic/phonological similarity during reading, it is also assumed that the reader has prior experience and knowledge regarding the frequency with which a given value is spelled with a given spelling, and will incorporate that knowledge during reading. As such, the communicative cost measure used here deviates slightly from similar prior work in the literature by further weighting the cost associated with communicating about a given value by the probability of that value being spelled with a given written form.

More precisely: The communicative cost $C(v)$ associated with communicating about a particular value v will be calculated according to (1):

$$C(v) = \sum_{g \in G} p(g|v) * \log_2\left(\frac{1}{R_g(v)}\right), \tag{1}$$

where G is the set of all written forms, and $R_g(v)$ is the similarity-based reader distribution:

$$R_g(v) \propto \sum_{w \in V} \text{sim}(v, w) * p(w|g), \tag{2}$$

where V is the set of all values (sound strings or morphemes, depending of the condition). More details on how similarity was calculated are given in appendix B.

The communicative cost $E[C]$ of the entire system is the expected communicative cost over written forms, as given in (3):

$$E[C] = \sum_{v} C(v)N(v). \tag{3}$$

The incorporation of need probability $N(v)$ (i.e., the probability that writers will need to communicate about value v) accounts for the aforementioned efficiency consideration that cultural needs will be reflected in that culture's languages.

Because of how it's being calculated, the range of possible communicative cost scores is partially a function of how many values there are and what the similarity relations are between those values, and will thus

differ for different languages. To make the communicative cost measure more comparison-friendly, it was normalized:

$$C = \frac{C_{\text{emp}} - C_{\text{min}}}{C_{\text{max}} - C_{\text{min}}}, \tag{4}$$

where C_{emp} is the empirical communicative cost score, and C_{max} and C_{min} are respectively the cost of the most uninformative and most informative possible systems for a given language and dataset. In line with the model of communication being used here, the least informative possible system is one which only has one form, which is used for all possible values, while the most informative system is one which assigns every unique value its own unique form. C_{max} and C_{min} can thus be calculated from the given data by either making all written forms identical or by assigning a unique written form to every value. Note that a normalized C score of 0 would mean that a system is optimally informative (relative to a given language and corpus).

Hypothetical Competitors

Recall that a system is efficient to the extent that it's as informative as possible given its complexity, and as simple as possible given its informativeness. As such, looking at a single system's informativeness and complexity in a vacuum doesn't necessarily reveal much about its efficiency; what needs to be shown is that the system is more efficient than other systems, plausible competitors which can, theoretically, convey the same messages. Examples of such plausible competitors include random baselines, variants of the real system that tweak certain parameters, or variants which are theoretically optimized relative to a particular goal. This strategy of hypothetical system generation and comparison is a well-established practice in the efficiency literature, and has been applied to a range of topics including word lengths (e.g., Pimentel, Nikkarinen, et al., 2021), word order (e.g., Futrell, Mahowald, and Gibson, 2015; Gildea and Temperley, 2010), pronouns (Denić, Steinert-Threlkeld, and Szymanik, 2020; Zaslavsky, Maldonado, and Culbertson, 2021), ambiguity (e.g., Trott and Bergen, 2020; 2022), grammatical marking (Mollica, Bacon, Zaslavsky, et al., 2021), quantifiers (Steinert-Threlkeld, 2021), and semantic typology (e.g., Karjus et al., 2021; Kemp and Regier, 2012; Zaslavsky, Kemp, Regier, and Tishby, 2018; Zaslavsky, Regier, Tishby, and Kemp, 2019).

This 'hypothetical competitor' methodology was adopted here as well: hypothetical variants of the target writing systems were created to serve as competitors against which the efficiency of the real systems could be evaluated. For all of these hypothetical systems, the only difference from the real system was the set of graphic forms and their mappings; everything else about the data was identical between systems. For

example, the frequency of a given value is the same for all systems, but the set of spellings used for that value can vary.

Most of the hypothetical systems used here involved randomly generating new graphic forms, relative to certain restrictions. Thousands of such systems were generated and compared against the real systems. For each generated system, a percentage (from 0% to 99%) of pre-existing mappings would be held constant; this was done so that the generated systems would span a range of plausibility, from variants that completely rewrite the existing system to variants which merely tweak a handful of the real system's mappings.

The suite of hypothetical systems is as follows:

Shuffled Systems (ShufS)

The existing mappings between graphic units and values are randomly shuffled. For example, if given the following 'real' system:

Graphic form	Pronunciation	Morpheme/meaning	Frequency
KA	ka	/kag/ "mouth"	1,300
KA	ka	/-ak/ "GENITIVE"	4,500
DU	gub	/gub/ "to carry"	1,000
DU	tum	/tum/ "to bring"	35

The possible shuffled systems would be any that randomly reorder the 'graphic form' column but keep all other columns fixed, such as:

Graphic form	Pronunciation	Morpheme/meaning	Frequency
KA	ka	/kag/ "mouth"	1,300
DU	ka	/-ak/ "GENITIVE"	4,500
DU	gub	/gub/ "to carry"	1,000
KA	tum	/tum/ "to bring"	35

If the real systems distribute their mappings in ways that facilitate communicative efficiency, then the expectation is that shuffling these mappings will disrupt both informativeness (similar values will no longer be more likely to share forms) and complexity (higher frequency values will no longer get assigned to simpler forms).

Graphotactically Generated (GGen)

New spellings are generated one character at a time using a weighted bigram language model with a 5 percent chance of ignoring the graphotactics and picking a character purely at random instead (this was done to

allow for the possibility of novel bigrams). While generating a spelling, the probability that the next character chosen x_n will be character $\hat{x}$ is:

$$p(x_n = \hat{x} \mid x_{n-1}) = (.05 * \frac{1}{|X|}) + (0.95 * p(x_{n-1} \mid \hat{x})),$$

where X is the set of all characters (including characters that mark the beginning or end of a spelling).

Phonographically Generated (PGen)

New spellings are generated for each item based on that item's phonological form. This is done by generating a new string of characters which can be read as the target phonological form based on those characters' extant character-sound mappings. The generation process is rerun for each graphic form-morpheme-phonological form triplet, meaning that homophones do not necessarily receive identical spellings. The PGen process also incorporates the empirical frequency of character-sound mappings, and when generating a new 'phonographic' spelling for a given sound string, it is more likely to pick characters that more often map to those sounds. This type of generated system was only run for the Sumerian analysis.

For example, Sumerian 𒈬 has among its possible readings /mu/ (as in *mu*$_2$ 'to grow') and /kiri/ (as in *kiri*$_6$ 'orchard'). In a PGen system, any phonological form that contains /mu/ or /kiri/ could be respelled such that the character 𒈬 is part of the new spelling. 𒈬 by itself would thus be a possible generated spelling for words such as *mu* 'name' or *kiri*$_3$ 'nose'. 𒈬 could also show up as part of the generated spelling for a word like *dumu* 'child' or *mušen* 'bird'.

In addition to the randomly generated systems described above, two 'curated' systems were included as points of comparison. These systems are, by design, highly phonographic, thus fitting a different typological profile than both the more logographic real systems and the generated systems (except for the pseudo-phonographic PGen systems). Because these systems are specifically catered to spelling-sound mappings, the expectation is that they will score better than the real systems for phonographic informativeness, but worse for semantic informativeness.

Syllabary Spellings (SylS)

All written tokens are respelled using a phonographically isomorphic syllabary (i.e., a set of a characters and mappings such that each character maps to exactly one extant syllable/mora and each extant syllable/mora maps to exactly one character). This results in all homophonous forms having identical spellings. The only polyvalence that

can remain is if a given morpheme maps to multiple pronunciations, in which case it will receive a different spelling for each of those pronunciations.

For Japanese, the syllabary was the existing katakana sub-system, i.e., the system was 'created' by simply using the katakana spelling for all morphemes. For Sumerian, a syllabary was created based on the character-sound mappings already present in the writing system. The process for creating this syllabary is detailed in appendix C.

Alphabetic System (AlphS)

All written tokens are respelled using the Roman alphabet, i.e., the romanization of that token's spoken form is treated as its spelling.

Results

Figures 1 (Sumerian) and 2 (Japanese) plot the extant writing systems against their respective hypothetical competitors. Table 4 shows the percentage of generated hypothetical systems that performed better than their respective real systems in terms of informativeness, complexity, and/or efficiency. For both Sumerian and Japanese, the real systems outperformed the hypothetical systems for at least one of, and more often both of, informativeness and complexity. A small percentage of hypothetical systems were more efficient than their respective real systems, with this being more common when informativeness was measured relative to spelling-sound mappings.

As expected, the more phonographic competitors were more phonographically informative. For Sumerian, this almost always came at the cost of system complexity: the hypothetical syllabary system, alphabetic system, and over 90% of the generated Sumerian PGen systems would require, on average, more strokes and more characters to write a given text as compared with the extant Sumerian writing system. The picture for Japanese is a bit more nuanced. If complexity is measured in characters, then the syllabary and alphabetic competitor systems follow the same pattern as for Sumerian, with the improved phonological informativeness offset by a substantial increase in complexity. However, if complexity is measured in strokes, then both of these systems are entirely more efficient than the real Japanese writing system. Also as expected, the phonographic variants of Japanese are worse with regards to semantic informativeness; however, it is noteworthy that the more phonographic hypothetical Sumerian systems are comparable to the real system for semantic informativeness.

For Japanese, the randomly shuffled competitors often demonstrated higher informativeness for sounds, but worse informativeness for mean-

TABLE 4. Percentage of generated hypothetical systems that were less complex, more informative, and/or more efficient than their respective real systems

	Sumerian				Japanese		
	ShufS	GGen	PGen	All	ShufS	GGen	All
Info. (Sem.)	65.33	0.00	13.64	27.56	0.24	0.00	0.11
Info. (Phon.)	1.20	0.00	96.73	26.39	85.18	0.00	41.37
Compl. (Stroke)	2.67	18.27	4.36	7.95	3.76	6.11	4.97
Compl. (Char.)	0.27	22.93	9.27	10.98	0.47	12.22	6.51
Effic (Sem. & Stroke)	0.13	0.00	0.73	0.24	0.00	0.00	0.00
Effic (Sem. & Char.)	0.00	0.00	1.64	0.44	0.00	0.00	0.00
Effic (Phon. & Stroke)	0.00	0.00	3.82	1.02	1.29	0.00	0.63
Effic (Phon. & Char.)	0.00	0.00	8.18	2.20	0.24	0.00	0.11

ings. Sumerian showed the opposite pattern, with the real system regularly outperforming the shuffled systems for phonological communicative cost, but performed worse for semantic communicative cost. It may seem surprising that the randomly shuffled systems regularly demonstrate higher informativeness than the real systems for any informativeness metric. One possible explanation is that the randomly shuffled systems result in less polyvalence overall. For example, the number of unique spelling types (N=2,281) is greater than the number of morpheme types (N=1,716) in Sumerian, with there being many high-effort but low-frequency spelling types. Because frequency doesn't factor in to the shuffling process, this may result in systems wherein high frequency morphemes get assigned unique but high-effort forms, which would account for the higher complexity and higher informativeness of these systems.

Discussion

Overall, the results are consistent with the efficiency hypothesis. With only a few exceptions, the plausible hypothetical competitor systems were less informative and/or more complex than the target systems. Before going in to how these results fit in to the literature more broadly, a few of the results bear some additional discussion.

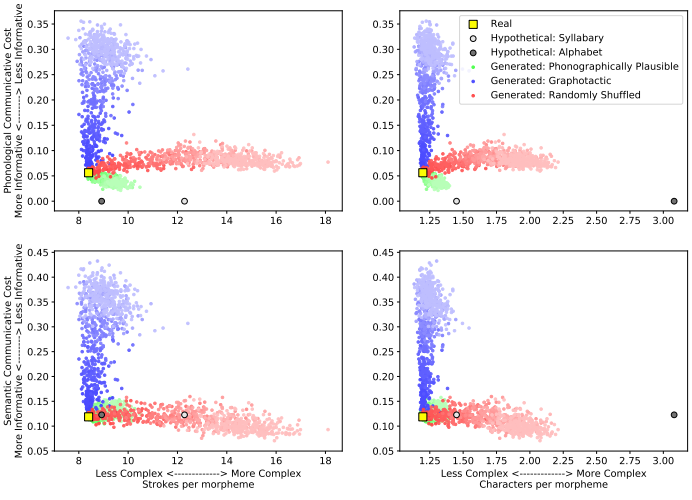

FIG. 1. Plot of Sumerian results. For the generated hypothetical systems, the darkness of the dots indicates the percentage of mappings that were fixed (i.e., identical to the real system): lighter dots represent systems with fewer fixed mappings.

More Efficient Systems

Even though a handful of generated systems did turn out to be more efficient than the real systems, these systems were only more efficient by a small margin. Furthermore, the only generated hypothetical systems which ever boasted better efficiency were those for which a high percentage of mappings were fixed during the generation process, meaning that they were mostly identical to the real systems. Nonetheless, the presence of any more efficient systems means that these results don't support the conclusion that the real systems are completely optimized for communicative efficiency, though the results are still consistent with the target systems being *near-optimized*. As discussed earlier, the argument here isn't that communicative efficiency is the only force shaping writing systems, so it is not a surprise, nor a challenge to the hypothesis, that the real systems are not as optimal as they possibly could be.

The biggest outliers among the hypothetical systems were the phonographic variants of written Japanese, which were the only hypothetical systems which were more efficient than a real system by a substantial margin (albeit only for one of the four informativeness-complexity pairings). It comes as no surprise that a variant of Japanese that abandons the use of logographic kanji and instead uses only katakana or rōmaji would

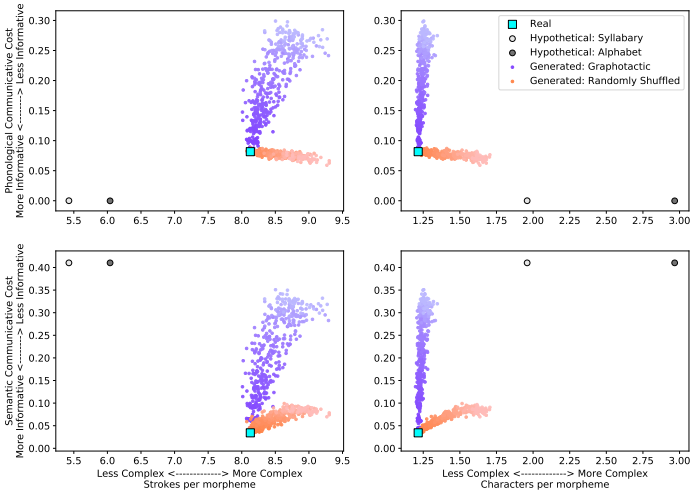

FIG. 2. Plot of Japanese results. For the generated hypothetical systems, the darkness of the dots indicates the percentage of mappings that were fixed (i.e., identical to the real system): lighter dots represent systems with fewer fixed mappings.

require fewer strokes and be more informative about sounds, and thus be more efficient relative to the goal of conveying phonological information. What is more surprising is just how much worse these systems would be regarding semantic informativeness, with the phonographic Japanese variants boasting worse semantic informativeness than even the randomly generated spellings. Furthermore, while katakana- and rōmaji-only systems would, by virtue of a reduced stroke count, require less motor effort (and likely be less visually complex), they would also require an increase in character tokens per text.

Logography, Phonography, and Linguistic Fit

Definitive conclusions about writing system typology and linguistic fit cannot be drawn based on a sample of only two languages/writing systems. What follows is a brief discussion of how the current results *could* be viewed as being consistent with the idea that spoken language morphology and phonology may influence what kinds of writing systems best enable efficiency communication.

Starting with Japanese, it is often argued/assumed that Japanese has too many homophones for a strictly sound-based writing system

to be effective, and that the retention of kanji is beneficial, if not necessary, for the purpose of homophone disambiguation (e.g., Backhouse (1984), Sampson (1985), and Smith (1996); Taylor and Taylor (1995, p. 323):323). The current results may support this view, to some extent: the phonographic Japanese variants were substantially less semantically informative than even the randomly generated, but still arguably logographic, competitors.

This pattern wasn't observed for Sumerian: the phonographic competitors and the real system had comparable semantic informativeness. Sumerian has fewer homophonous morphemes than Japanese[11], but the results suggest that Sumerian's logographic writing system was efficient. This in turn suggests that a high number of homophones is not the only factor that would push a writing system towards greater logography.

More broadly, the results for the phonographic competitors paint an interesting picture relative to the presumed merits of phonography over logography. The expectation that the phonographic variants would be more phonographically informative is indeed borne out. However, in many cases this came at the cost of both complexity *and* semantic informativeness: one step forward, but two steps back. One interpretation of this result is that, at least for Sumerian and Japanese, a shift towards phonography may well result in the systems actually being less efficient overall. (Of course, a shift towards phonography may yield benefits for factors not considered in the current analysis, such as rate of acquisition.) If nothing else, the phonographic competitors do not seem to be *more* efficient than the original logographic systems, which runs counter to argument that logography is inherently disadvantageous.

Modality and Efficiency

As mentioned earlier, while the body of evidence in favor of the efficiency hypothesis continues to grow, much of this evidence is either based on spoken language or is modality-agnostic. The current study, by contrast, very directly focuses in on writing systems, specifically two writing systems which are not direct 'mirrors' of their respective spoken languages, especially with regards to the instantiation of ambiguity. In other words, the observed efficiency of these systems, ambiguity and all, is not merely 'inherited' from their spoken languages.

The results not only support the hypothesis that writing systems are shaped by functional communicative pressures, but by extension, and in conjunction with the existing efficiency literature, further support the

11. In the current datasets roughly 90% of underlying phonological forms mapped to two or more morphemes for Japanese, but only around 50% of URs were homophonous for Sumerian (at least based on modern understanding of Sumerian phonology).

view that human linguistic communication in general, independent of modality, is shaped by such pressures. However, the question of whether modality affects the realization of efficiency remains open.

Future Considerations

The study presented here only addresses a small piece of the broader question of whether efficiency pressures shape writing systems and written language. To close, it may be worth briefly discussing some of the ways in which this kind of study can be improved in future research.

Optimization and the Space of Possible Systems

One consequence of considering an entire writing system/written language is that the space of all *possible* hypothetical variants is intractably large (and, depending on limitations, theoretically infinite). In comparison with some past efficiency studies which narrowed in on specific semantic domains, this means that it's not easy to tell how a given writing systems stacks up complexity-wise in the space of all possible systems. As such, future work may benefit from establishing plausible optimized competitor systems (in a similar vein as the syllabary/alphabetic systems used here) which can serve as benchmarks of comparison. The possibility space being language-specific also means that, again unlike other work in the efficiency literature, the two target writing systems couldn't be directly compared against each other in the same space. Establishing a language-neutral space is a challenge, since phonology and lexicons are so language-dependent, but doing so would allow for more direct comparisons of writing systems.

Complexity

While the two types of complexity used here are certainly relevant for writing systems and constitute a strong starting point, an analysis of writing system efficiency would benefit from considering a wider array of possible types of complexity and effort, such as inventory size, learnability, and visual complexity. Perhaps chief among these is the size of the inventory of characters (or other graphic units), and by extension the costs associated with learning and retaining a system. For example, while the hypothetical syllabary-only systems considered in the current study tended towards greater motor effort and/or material costs, they also presumably resulted in systems which would require learning a much smaller number of distinct characters.

Informativeness

Given the significant role that context has been argued to play in the efficiency of ambiguity (Piantadosi, Tily, and Gibson, 2012), informativeness measures may benefit from incorporating contextual information, which wasn't done in the current study. Previous work (Hermalin and Regier, 2019) has already found evidence that Sumerian character ambiguity is greatly reduced by even a small amount of additional context, but the overall in-context efficiency of written Sumerian, as well as written Japanese, remains to be tested.

Another potential issue with the current implementation of informativeness is that the similarity-based reader distribution always prefers that more similar referents have the same form. However, as argued in Brochhagen and Boleda (2022), it may be more advantageous for two meanings to use different forms if those two meanings are *too* similar. As such, similarity-based informativeness measures used in future work may benefit from assuming a less monotonic relationship between similarity and informativeness. The task of quantifying semantics, especially for ancient or low-resource languages, is itself another challenge for research in this vein (see appendix B for details, and critiques, of how the current study handled this issue).

Conclusion

The results reported here provide evidence that even highly logographic and polyvalent writing systems are communicatively efficient. This suggests that neither logography nor ambiguity are inherently disadvantageous for writing systems, and their presence may even serve to enhance efficiency.

As discussed earlier, the idea of writing systems being shaped by efficiency pressures is not at all new, but more data-driven, quantitative approaches to measuring the communicative efficiency of writing systems have remained elusive. Regardless of whether or not the approach used here proves to be the best option, it is the hope of the current author that researchers interested in writing systems and functional cognitive pressures on human communication will continue to consider how matters of writing system efficiency, development, and linguistic fit can be measured in quantitative, theoretically grounded ways.

A. Glossary and Terminology

The following covers how certain writing systems-related terminology are defined in this paper. Many of the definitions used here are adapted

from Neef (2015) and Weingarten (2011). The definitions used here all intended to be somewhat flexible.

- *Character*: The primary meaning/information-bearing unit of written form in a writing system. Characters typically (but not always): occupy a certain amount of space, can map to some stable portion of linguistic information (such as a phoneme or morpheme), are treated as distinct units in sociological contexts, and are stable in the sense that they can 'stand on their own'.
- *Spelling*: A sequence of one or more characters which can be cleanly mapped to a morpheme or sound sequence.
- *Script*: A set of written forms, (characters, diacritics, etc.) which are used by a writing system.
- *Written Language*: Language or linguistic communication for which the modality is writing.
- *Non-written Language*: Language or linguistic communication that takes place in any modality other than writing. (Note that this term is *not* being used with the meaning of "a language that doesn't currently have a writing system or written counterpart".)
- *Writing System*: A system which maps written forms, composed of components drawn from a script, to units of linguistic information, ultimately to meaning-bearing units such as morphemes. More generally, a writing system consists of a script, a language, and a mapping from one to the other.
- *Phonography and Logography*: Properties and uses of a writing system that primarily pertain to sounds and spoken forms will be considered phonographic. Properties and uses of a writing system that pertain to semantics and/or lexical identity, or which involve a single character mapping to a morpheme or entire word, will fall under the header of logography. Terms such as *semantographic* and *morphographic* may be used to more precisely distinguish between different uses of *logographic*.
- *Syllabary*: A writing system in which a character can map to a syllable or mora.

B. Methods—Similarity

Phonological similarity was calculated as a function of phonological distance. Phonological distance was calculated as feature-weighted Levenshtein edit distance, using methods adapted from Abdullah et al. (2021) and Fontan et al. (2016). The distance between two strings *a* and *b* is based on how many operations (deletion, insertion, or replacement) are needed to turn *a* into *b* and vice versa, with the cost associated with replacement being based on the number of phonological features that are identical between the two sounds involved. More precisely, the similarity between two individual sounds was treated as the number of phono-

logical features for which they had the same value divided by the total number of features which were applicable to both sounds; this results in a distance score between 0 and 1. Feature values for each phone were based on the featural specifications provided by PHOIBLE 2.0 (Moran and McCloy, 2019). These feature values were constrained by which features were distinctive for the target language (for example, since Japanese doesn't have any ejective sounds, features that distinguish between ejective and pulmonic sounds were not included for the feature values of phones in Japanese).

The phonological similarity between two sound strings a and b was handled as follows:

1. Both strings were syllabified. Within each syllable, each segment was marked as either being part of the onset, nucleus, or coda.
2. For each of the syllables $\sigma_{a1} \cdots \sigma_{an}$ of a, the distance was calculated between that syllable and each of the syllables $\sigma_{b1} \cdots \sigma_{bn}$ of b.
3. a and b were aligned so as to minimize the total distance between aligned syllables. If a and b differed in number of syllables, the shorter form was padded with empty syllables.
4. The total distance between a and b was then calculated as the sum of featural distances between aligned syllables.

For example, if the two target strings were /ugula/ and /udu/, the process would: syllabify the strings into /u.gu.la/ and /u.du/; get the distance between each syllable of the first string and each syllable of the second string (i.e., dist(u, u), dist(u, du), dist(gu, u), dist(gu, du), etc.); align the two syllabified strings so as to minimize the featural distance between aligned syllables, resulting in the alignment $\begin{matrix} \text{u} & \text{gu} & \text{la} \\ \text{u} & \text{du} & \emptyset \end{matrix}$; then sum the featural distance of each aligned syllable pair.

For semantic similarity, morphemes were converted to numerical vectors, with the similarity between two morphemes being a function of the cosine distance between their respective vectors. The pretrained Sentence-BERT model (Reimers and Gurevych, 2019) was used to convert morphemes to 384-dimensional vector embeddings. Because this model is trained on English, the inputs were English definitions of the target morphemes. For Sumerian, the definitions used were those given by ePSD2's online dictionary (`https://oracc.museum.upenn.edu/epsd2/sux/`). For Japanese, definitions were retrieved using the Jamdict Python library (`https://jamdict.readthedocs.io/en/latest/`) and, for certain Sino-Japanese morphemes, kanjidatabase.com's *jukugo* database (Tamaoka, Makioka, Sanders, and Verdonschot, 2017)[12].

12. This is an imperfect method of quantifying semantics, for a few reasons. The most prominent flaw is that it assumes that a model trained on modern English sen-

To strike a balance between how similar to morphemes are in absolute terms and how similar they are relative to other morphemes, the semantic similarity between two morphemes $\text{sim}(m_1, m_2)$ was calculated as the mean of two similarity measures, sim_{BERT} and sim_{rank}. The sim_{BERT} measure is simply the BERT-based similarity score, as described above. The sim_{rank} measure is based on the rank BERT similarity between the two morphemes, where the similarity rank of morpheme m_2 relative to morpheme m_1 equals 1 if m_2 is the most similar morpheme to m_1, is 2 if m_2 is the second most similar morpheme to m_1, etc. To keep the sim measure symmetric (such that $\text{sim}(m_1, m_2) = \text{sim}(m_2, m_1)$), the sim_{rank} measure between two morphemes was taken as the average similarity rank $\text{rank}_{\text{mean}}(m_1, m_2)$ between the two. This score was then normalized, with similar pairs scoring closer to 1 and less similar pairs scoring closer to 0.

C. Hypothetical Sumerian Syllabary

The hypothetical Sumerian syllabary used in the current study was created as follows: For each unique syllable (going in order from most frequent to least frequent), get the list of individual characters that map to that syllable at least once in the corpus and which haven't already been assigned by this syllabary-generation process. Of those characters, choose the one that most often maps to the target syllable and assign it to the target syllable. That character now has the syllabary value of

tences and modern English texts will accurately capture meaning distinctions and associations for morphemes in other languages. This semantics approach also takes at face value the definitions provided by the respective translation dictionaries, but even thorough dictionary definitions can't capture the full extent of a word's meanings. As a result, it is assuredly the case that certain semantic similarities or associations which are present in the target languages will not be captured by the translation-based BERT embeddings, and certain associations which are present only for English will be falsely applied to Sumerian and Japanese. The reasons behind adopting this flawed approach are mostly based in convenience and data availability. Training a new language model on the comparatively small (and genre-imbalanced) Sumerian corpus carries risks of its own; furthermore, modern lexicography of Sumerian word meanings may include information which, by chance, wouldn't be captured in the Ur3Admin Corpus, meaning that translated definitions may actually be more informative than a corpus-trained model, at least in some ways. Overall, the use of translated definitions is somewhat of a necessity for Sumerian. The decision to use the same strategy for Japanese, rather than use existing Japanese-trained models, was made so as to make the treatment of both languages as similar as possible, but it would assuredly be an improvement to approach Japanese semantics using Japanese data. Nonetheless, it is the opinion of the author that, for all its flaws, using English translations and an English-trained model does still capture accurate information about lexical semantics, and while the approach is certainly imperfect, it is not invalid.

that syllable, and cannot be assigned additional syllable values. If there are no available characters for a syllable, either because they have no extant single-character spellings or because all of their candidate characters had already been assigned other mappings, then that syllable is temporarily skipped. Once every syllable has either been assigned or skipped in this way, the not-yet-assigned syllables are mapped as follows: for each remaining syllable, choose a random not-yet-assigned character which has among its syllable values a syllable that has the same onset+nucleus or the same rime as the target syllable, and assign that character to the target syllable. For example, if the target syllable was /bab/, any character that has /ba/ or /ab/ among its values would be a viable candidate. This process still yields a small number of C_1VC_2 syllables which don't get assigned. These remaining syllables get treated as if they were C_1V-VC_2 sequences, meaning that they will be spelled as the character that's been assigned to C_1V followed by the character that's been assigned to VC_2. (The use of this C_1V-VC_2 strategy is justified by it being an actual spelling strategy that was used for writing Sumerian and Akkadian.)

This results in a hypothetical Sumerian syllabary for which: every character has exactly one value, and that value is a syllable; and every unique syllable (that occurs in the corpus) has exactly one spelling[13]. In most cases, this spelling is a single character, with the only exceptions being a handful of C_1VC_2 syllables syllables which get 're-analyzed' as C_1V-VC_2 sequences, as described above. Note that this system is not optimized for anything other than phonological informativeness: for example, when assigning values to characters, the most common syllables weren't automatically paired with the lowest complexity characters. Instead, this generation process was intended to at least approximately result in a system which could plausibly have arisen from the actual writing system used for writing Ur III Sumerian, and can be used as a stand-in for the 'what if' scenario where the actual Sumerian writing system developed into a dedicated syllabary.

13. While this hypothetical system resembles the generated PGen systems described above insofar as new spellings are created based on existing character-sound mappings, the Sumerian SylS system differs from the PGen systems in some key ways: the SylS system is (almost) completely unambiguous, at the level of both morpheme spellings (from the writer's perspective) and individual syllable spellings (for both writer and reader), whereas the PGen systems maintained varying levels of polyvalence; the SylS system is purely syllabographic, while the PGen systems included mappings such that a single character could map to a string of two or more syllables; and the SylS system isn't randomly generated, while the PGen systems were.

References

Abdullah, Badr M. et al. (2021). "Do acoustic word embeddings capture phonological similarity? An empirical study". arXiv:2106.08686.

Aylett, Matthew and Alice Turk (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech". In: *Language and speech* 47.1, pp. 31–56.

Backhouse, A. E. (1984). "Aspects of the graphological structure of Japanese". In: *Visible Language* 18.3, p. 219.

Baines, John (2004). "The earliest Egyptian writing: development, context, purpose". In: *The first writing: Script invention as history and process*. Cambridge University Press Cambridge, pp. 150–189.

Baroni, Antonio (2011). "Alphabetic vs. non-alphabetic writing: Linguistic fit and natural tendencies". In: *Rivista di Linguistica* 23.2, pp. 127–159.

Berg, Kristian and Mark Aronoff (2017). "Self-organization in the spelling of English suffixes: The emergence of culture out of anarchy". In: *Language*, pp. 37–64.

Black, Jeremy A. (1990). "The alleged "extra" phonemes of Sumerian". In: *Revue d'Assyriologie et d'archéologie orientale* 84.2, pp. 107–118.

Boltz, William G. (2011). "Literacy and the Emergence of Writing in China". In: *Writing and Literacy in Early China: Studies from the Columbia Early China Seminar*. University of Washington Press Seattle, pp. 51–84.

Borgwaldt, Susanne R., Frauke M. Hellwig, and Annette MB de Groot (2004). "Word-initial entropy in five languages: Letter to sound, and sound to letter". In: *Written Language & Literacy* 7.2, pp. 165–184.

Borleffs, Elisabeth et al. (2017). "Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review". In: *Reading and writing* 30, pp. 1617–1638.

Brochhagen, Thomas (2020). "Signalling under uncertainty: Interpretative alignment without a common prior". In: *The British Journal for the Philosophy of Science*.

Brochhagen, Thomas and Gemma Boleda (2022). "When do languages use the same word for different meanings? The Goldilocks principle in colexification". In: *Cognition* 226, p. 105179.

Cahill, Michael and Elke Karan (2008). "Factors in designing effective orthographies for unwritten languages". In: *SIL International*.

Caplan, Spencer, Jordan Kodner, and Charles Yang (2020). "Miller's monkey updated: Communicative efficiency and the statistics of words in natural language". In: *Cognition* 205, p. 104466.

Chang, Li-Yun, Yen-Chi Chen, and Charles A. Perfetti (2018). "GraphCom: A multidimensional measure of graphic complexity applied to 131 written languages". In: *Behavior research methods* 50, pp. 427–449.

Changizi, Mark A. and Shinsuke Shimojo (2005). "Character complexity and redundancy in writing systems over human history". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1560, pp. 267–275.
Changizi, Mark A. et al. (2006). "The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes". In: *The American Naturalist* 167.5, E117–E139.
Chen, Ping (1999). *Modern Chinese: History and Sociolinguistics.* ERIC.
Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*.
Civil, Miguel (1973). "The Sumerian writing system: Some problems". In: *Orientalia* 42, pp. 21–34.
———— (2004). "The Series DIRI = (w)atru". In: *Materials for the Sumerian Lexicon* XV.
Cohen Priva, Uriel (2017). "Informativity and the actuation of lenition". In: *Language*, pp. 569–597.
Cooper, Jerrold S. (2004). "Babylonian beginnings: the origin of the cuneiform writing system in comparative perspective". In: *The first writing: Script invention as history and process*. Cambridge University Press Cambridge, pp. 71–99.
Coulmas, Florian (2003). *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press.
———— (2009). "Evaluating merit—the evolution of writing reconsidered". In: *Writing systems research* 1.1, pp. 5–17.
———— (2018). "Writing and Literacy in Modern Japan". In: *The Cambridge Handbook of Japanese Linguistics*. Ed. by Yoko Hasegawa. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, pp. 114–132.
Coupé, Christophe et al. (2019). "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche". In: *Science advances* 5.9, eaaw2594.
Cover, T. M. and Joy A. Thomas (1991). *Elements of information theory*. eng. Wiley series in telecommunications. New York: Wiley. ISBN: 0471062596.
Daniels, Peter T. (1990). "Fundamentals of grammatology". In: *Journal of the American Oriental Society*, pp. 727–731.
———— (1992). "The syllabic origin of writing and the segmental origin of the alphabet". In: *The linguistics of literacy* 21, pp. 83–110.
———— (2017). "Writing systems". In: *The handbook of linguistics*. Wiley Online Library, pp. 75–94.
Daniels, Peter T. and William Bright (1996). *The world's writing systems*. Oxford University Press on Demand.
Daniels, Peter T. and David L. Share (2018). "Writing system variation and its consequences for reading and dyslexia". In: *Scientific Studies of Reading* 22.1, pp. 101–116.
Dautriche, Isabelle et al. (2017). "Words cluster phonetically beyond phonotactic regularities". In: *Cognition* 163, pp. 128–145.

de Voogt, Alex (2024). "The Evolution of Writing Systems: An Introduction". In: *Oxford Handbook of Human Symbolic Evolution*. Oxford University Press.

DeFrancis, John (1989). *Visible speech: The diverse oneness of writing systems*. University of Hawaii Press.

———— (1996). "How efficient is the Chinese writing system?" In: *Visible Language* 30.1, p. 06.

Denić, Milica, Shane Steinert-Threlkeld, and Jakub Szymanik (2020). "Complexity/informativeness trade-off in the domain of indefinite pronouns". In: *Semantics and linguistic theory*, pp. 166–184.

Devlin, Jacob et al. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding". preprint arXiv:1810.04805.

Diringer, David (1962). *Writing*. Frederick A. Praeger.

Dye, Melody et al. (2017). "A functional theory of gender paradigms". In: *Perspectives on morphological organization*. Brill, pp. 212–239.

———— (2018). "Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication". In: *Topics in cognitive science* 10.1, pp. 209–224.

Edzard, Dietz Otto (2003). *Sumerian grammar*. Vol. 71. Brill.

Fedzechkina, Maryia, T. Florian Jaeger, and Elissa L. Newport (2011). "Functional biases in language learning: Evidence from word order and case-marking interaction". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 33. 33.

Fedzechkina, Maryia, Elissa L. Newport, and T. Florian Jaeger (2017). "Balancing effort and information transmission during language acquisition: Evidence from word order and case marking". In: *Cognitive science* 41.2, pp. 416–446.

Fedzechkina, Masha, Lucy Hall Hartley, and Gareth Roberts (2023). "Social biases can lead to less communicatively efficient languages". In: *Language Acquisition* 30.3-4, pp. 230–255.

Fontan, Lionel et al. (2016). "Using phonologically weighted Levenshtein distances for the prediction of microscopic intelligibility". In: *Annual conference Interspeech (INTERSPEECH 2016)*, pp. 650–654.

Fortson, Benjamin W. (2023). "Systems and Idiosyncrasies". In: *The Cambridge Handbook of Historical Orthography*. Cambridge University Press, pp. 183–203.

Foxvog, Daniel A. (2014). "Introduction to Sumerian grammar". In: *TCS* 1.60, p. 3.

Frellesvig, Bjarke (2010). *A History of the Japanese Language*. Cambridge University Press.

Frost, Ram (2012). "Towards a universal model of reading". In: *Behavioral and brain sciences* 35.5, pp. 263–279.

Futrell, Richard, Kyle Mahowald, and Edward Gibson (2015). "Large-scale evidence of dependency length minimization in 37 languages". In: *Proceedings of the National Academy of Sciences* 112.33, pp. 10336–10341.

Gahl, Susanne, Yao Yao, and Keith Johnson (2012). "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech". In: *Journal of memory and language* 66.4, pp. 789–806.
Garrod, Simon et al. (2007). "Foundations of representation: where might graphical symbol systems come from?" In: *Cognitive science* 31.6, pp. 961–987.
Gelb, Ignace J. (1963). *A study of writing*. University of Chicago Press.
Gibson, Edward et al. (2019). "How efficiency shapes human language". In: *Trends in Cognitive Sciences* 23.5, pp. 389–407.
Gildea, Daniel and David Temperley (2010). "Do grammars minimize dependency length?" In: *Cognitive Science* 34.2, pp. 286–310.
Halliday, M. A. K. (2010). "Ideas About Language (1977)". eng. In: *On Language and Linguistics*. Continuum International Publishing Group, pp. 92–115. ISBN: 1474211933.
Han, Simon J. et al. (2022). "Simplification is not dominant in the evolution of Chinese characters". In: *Open Mind* 6, pp. 264–279.
Handel, Zev (2013). "Can a logographic script be simplified? Lessons from the 20th century Chinese writing reform informed by recent psycholinguistic research". In: *Scr. Theol* 5, pp. 21–66.
———— (2019). *Sinography: The borrowing and adaptation of the Chinese script*. Brill.
Hannas, William C. (1997). *Asia's orthographic dilemma*. University of Hawaii Press.
Hermalin, Noah (2023). "A mutual information-based approach to quantifying logography in Japanese and Sumerian". In: *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pp. 105–110.
Hermalin, Noah and Terry Regier (2019). "Efficient use of ambiguity in an early writing system: Evidence from Sumerian cuneiform." In: *CogSci*, pp. 422–427.
Honda, Keisuke (2007). "Homography-induced ambiguity in Japanese kanji and the lexical disambiguating function of okurigana". In: *Gengogaku Ronsō* 26, pp. 17–42.
———— (2008). "Okurigana: An Analysis from the Viewpoint of Decoding". In: 筑波応用言語学研究 *[Journal of Applied Linguistics at Tsukuba]* 15, pp. 87–100.
Houston, Stephen D. (2012a). "The Shape of Script—Views from the Middle". In: *The shape of script: How and why writing systems change*. Ed. by Stephen D. Houston. School for Advanced Research Press. Santa Fe, pp. xiii–xxiii.
———— ed. (2012b). *The shape of script: How and why writing systems change*. School for Advanced Research Press.
Howes, Davis (1968). "Zipf's law and Miller's random-monkey model". In: *The American Journal of Psychology* 81.2, pp. 269–272.

Jaeger, T. and Roger Levy (2006). "Speakers optimize information density through syntactic reduction". In: *Advances in neural information processing systems* 19.

Jaeger, T. Florian (2010). "Redundancy and reduction: Speakers manage syntactic information density". In: *Cognitive psychology* 61.1, pp. 23–62.

Jagersma, Bram (2010). "A descriptive grammar of Sumerian". PhD thesis. Leiden University.

Joyce, Terry and Dimitrios Meletis (2021). "Alternative criteria for writing system typology: Cross-linguistic observations from the German and Japanese writing systems". In: *Zeitschrift für Sprachwissenschaft* 40.3, pp. 257–277.

Juba, Brendan et al. (2011). *Compression without a common prior: an information-theoretic justification for ambiguity in language*. Institute for Theoretical Computer Science.

Justeson, John S. (1978). "Mayan scribal practice in the Classic period: A test-case of an explanatory approach to the study of writing systems". PhD thesis. Stanford University.

Kageyama, Taro and Michiaki Saito (2016). "Vocabulary strata and word formation processes". In: *Handbook of Japanese lexicon and word formation*. Mouton de Gruyter Berlin, pp. 11–50.

Karjus, Andres et al. (2021). "Conceptual similarity and communicative need shape colexification: An experimental study". In: *Cognitive Science* 45.9, e13035.

Katz, Leonard and Ram Frost (1992). "The reading process is different for different orthographies: The orthographic depth hypothesis". In: *Advances in psychology*. Vol. 94. Elsevier, pp. 67–84.

Kelly, Piers et al. (2021). "The predictable evolution of letter shapes: An emergent script of West Africa recapitulates historical change in writing systems". In: *Current Anthropology* 62.6, pp. 669–691.

Kemp, Charles and Terry Regier (2012). "Kinship categories across languages reflect general communicative principles". In: *Science* 336.6084, pp. 1049–1054.

Kemp, Charles, Yang Xu, and Terry Regier (2018). "Semantic typology and efficient communication". In: *Annual Review of Linguistics* 4, pp. 109–128.

Koshevoy, Alexey, Helena Miton, and Olivier Morin (2023). "Zipf's law of abbreviation holds for individual characters across a broad range of writing systems". In: *Cognition* 238, p. 105527.

Kuperman, Victor and Raymond Bertram (2013). "Moving spaces: Spelling alternation in English noun-noun compounds". In: *Language and Cognitive processes* 28.7, pp. 939–966.

Kurumada, Chigusa and T. Florian Jaeger (2015). "Communicative efficiency in language production: Optional case-marking in Japanese". In: *Journal of Memory and Language* 83, pp. 152–178.

Lee, Hanjung (2009). "Quantitative variation in Korean case ellipsis: implications for case theory". In: *Differential subject marking*. Springer, pp. 41–61.

Levshina, Natalia (2021). "Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations". In: *Frontiers in Psychology* 12, p. 648200.

——— (2022). *Communicative efficiency*. Cambridge University Press.

Maekawa, Kikuo et al. (2014). "Balanced corpus of contemporary written Japanese". In: *Language resources and evaluation* 48, pp. 345–371.

Mahowald, Kyle, Isabelle Dautriche, et al. (2022). "Efficient Communication and The Organization of The Lexicon". eng. In: *The Oxford Handbook of the Mental Lexicon*. Oxford Handbooks. Oxford University Press. ISBN: 9780198845003.

Mahowald, Kyle et al. (2013). "Info/information theory: Speakers choose shorter words in predictive contexts". In: *Cognition* 126.2, pp. 313–318.

Majid, Asifa et al. (2018). "Differential coding of perception in the world's languages". In: *Proceedings of the National Academy of Sciences* 115.45, pp. 11369–11376.

Marjou, Xavier (2019). "Oteann: Estimating the transparency of orthographies with an artificial neural network". arXiv:1912.13321.

Mattingly, Ignatius G. (1985). "Did orthographies evolve?" In: *Remedial and special education* 6.6, pp. 18–23.

Meister, Clara et al. (2021). "Revisiting the uniform information density hypothesis". preprint arXiv:2109.11635.

Meletis, Dimitrios (2019). "Naturalness in scripts and writing systems: Outlining a Natural Grapholinguistics". PhD thesis. University of Graz.

Michalowski, Piotr (2004). "Sumerian". In: *The Cambridge Encyclopedia of the World's Ancient Languages*, pp. 19–59.

——— (2020). "Sumerian". In: *A companion to ancient Near Eastern languages*, pp. 83–105.

Miton, Helena and Olivier Morin (2021). "Graphic complexity in writing systems". In: *Cognition* 214, p. 104771.

Mollica, Francis, Geoff Bacon, Yang Xu, et al. (2020). "Grammatical marking and the tradeoff between code length and informativeness." In: *CogSci*.

Mollica, Francis, Geoff Bacon, Noga Zaslavsky, et al. (2021). "The forms and meanings of grammatical markers support efficient communication". In: *Proceedings of the National Academy of Sciences* 118.49, e2025993118.

Moran, Steven and Daniel McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. URL: https://phoible.org/.

Morin, Olivier (2018). "Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy". In: *Cognitive Science* 42.2, pp. 664–677.
Morin, Olivier and Alexey Koshevoy (2024). "A Cultural Evolutionary Model for the Law of Abbreviation". In: *Topics in cognitive science*.
Morita, Aiko and Katsuo Tamaoka (2002). "Semantic involvement in the lexical and sentence processing of Japanese Kanji". In: *Brain and language* 82.1, pp. 54–64.
Nagano, Akiko and Masaharu Shimada (2014). "Morphological theory and orthography: Kanji as a representation of lexemes". In: *Journal of Linguistics* 50.2, pp. 323–364.
Neef, Martin (2015). "Writing systems as modular objects: proposals for theory design in grapholinguistics". In: *Open Linguistics* 1.1.
Osterkamp, Sven and Gordian Schreiber (2021). "<The>e ubi<qu>ity of polygra<ph>y and its significan<ce> for <th>e typology of <wr>iti<ng> systems". In: *Written Language & Literacy* 24.2, pp. 171–197.
Peloquin, Benjamin N., Noah D. Goodman, and Michael C. Frank (2020). "The interactions of rational, pragmatic agents lead to efficient language structure and use". In: *Topics in Cognitive Science* 12.1, pp. 433–445.
Piantadosi, Steven T., Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529.
———— (2012). "The communicative function of ambiguity in language". In: *Cognition* 122.3, pp. 280–291.
Pimentel, Tiago, Clara Meister, et al. (2023). "Revisiting the Optimality of Word Lengths". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2240–2255.
Pimentel, Tiago, Irene Nikkarinen, et al. (2021). "How (non-) optimal is the lexicon?" preprint arXiv:2104.14279.
Pimentel, Tiago et al. (2020). "Speakers fill lexical semantic gaps with context". preprint arXiv:2010.02172.
Powlison, Paul S. (1968). "Bases for formulating an efficient orthography". In: *The Bible Translator* 19.2, pp. 74–91.
Protopapas, Athanassios and Eleni L. Vlahou (2009). "A comparative quantitative analysis of Greek orthographic transparency". In: *Behavior research methods* 41, pp. 991–1008.
Pulgram, Ernst (1976). "The typologies of writing systems". In: *Writing without letters*. Manchester University Press, Manchester, pp. 1–28.
Regier, Terry, Alexandra Carstensen, and Charles Kemp (2016). "Languages support efficient communication about the environment: Words for snow revisited". In: *PloS one* 11.4, e0151138.
Regier, Terry, Charles Kemp, and Paul Kay (2015). "Word meanings across languages support efficient communication". In: *The handbook of language emergence*, pp. 237–263.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". arXiv:1908.10084.
Rogers, Henry (1995). "Optimal orthographies". In: *Scripts and literacy: Reading and learning to read alphabets, syllabaries and characters*. Springer, pp. 31–43.
Salomon, Richard (2012). "Some principles and patterns of script change". In: *The shape of script: How and why writing systems change*, pp. 119–133.
Sampson, Geoffrey (1985). "Writing Systems". In: *London, UK: Hutchinson*.
———— (2015). "Writing systems: methods for recording language". In: *The Routledge handbook of linguistics*. Routledge, pp. 47–61.
———— (2016). "Typology and the study of writing systems". In: *Linguistic Typology* 20.3, pp. 561–567.
———— (2018). "From phonemic spelling to distinctive spelling". In: *Written Language & Literacy* 21.1, pp. 3–25.
Seeley, Christopher (2000). *A history of writing in Japan*. University of Hawaii Press.
Seidenberg, Mark S. (2011). "Reading in different writing systems: One architecture, multiple solutions." In: *Dyslexia across Languages: Orthography and the Brain-Gene-Behavior Link.* Paul H. Brookes Publishing, pp. 146–168.
———— (2012). "Writing systems: Not optimal, but good enough". In: *Behavioral and Brain Sciences* 35.5, p. 305.
Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, pp. 379–423.
Share, David L. (2014). "Alphabetism in reading science". In: *Frontiers in psychology* 5, p. 752.
Siegelman, Noam, 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". In: *Behavior research methods* 52, pp. 1292–1312.
Siegelman, Noam, Jay G. Rueckl, et al. (2022). "Quantifying the regularities between orthography and semantics and their impact on group-and individual-level behavior." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 48.6, p. 839.
Smith, Janet S. (Shibatani) (1996). "Japanese Writing". In: *The world's writing systems*. Oxford University Press, pp. 209–217.
Sproat, Richard (2023). *Symbols: An evolutionary history from the stone age to the future*. Springer Nature.
Sproat, Richard and Alexander Gutkin (2021). "The taxonomy of writing systems: How to measure how logographic a system is". In: *Computational Linguistics* 47.3, pp. 477–528.
Steinert-Threlkeld, Shane (2021). "Quantifiers in natural language: Efficient communication and degrees of semantic universals". In: *Entropy* 23.10, p. 1335.

Tamaoka, Katsuo et al. (2017). "www.kanjidatabase.com: a new interactive online database for psychological and linguistic research on Japanese kanji and their compound words". In: *Psychological research* 81, pp. 696–708.

Taylor, Insup and M. Martin Taylor (1995). *Writing and literacy in Chinese, Korean, and Japanese*. Studies in written language and literacy, v. 3. Amsterdam ; John Benjamins Pub. Co. ISBN: 9027217947.

Taylor, Isaac (1883). *The alphabet: an account of the origin and development of letters.* eng. London: K. Paul, Trench & co.

Tinney, Steve and Eleanor Robson (2014). *ORACC: The Open Richly Annotated Cuneiform Corpus*. URL: http://oracc.museum.upenn.edu.

Trigger, Bruce G. (1998). "Writing systems: A case study in cultural evolution". In: *Norwegian archaeological review* 31.1, pp. 39–62.

Trott, Sean and Benjamin Bergen (2020). "Why do human languages have homophones?" In: *Cognition* 205, p. 104449.

———— (2022). "Languages are efficient, but for whom?" In: *Cognition* 225, p. 105094.

Unger, J. Marshall (2011). "What linguistic units do Chinese characters represent?" In: *Written Language & Literacy* 14.2, pp. 293–302.

———— (2014). "Empirical evidence and the typology of writing systems: a response to Handel". In: *SCRIPTA, Volume 6*, pp. 75–95.

Unger, J. Marshall and John DeFrancis (1995). "Logographic and semasiographic writing systems: A critique of Sampson's classification". In: *Scripts and literacy: Reading and learning to read alphabets, syllabaries and characters*. Springer, pp. 45–58.

Vachek, Josef (1973). *Written language: General problems and problems of English*. The Hague: Mouton.

Van den Bosch, Antal et al. (1994). "Measuring the complexity of writing systems". In: *Journal of Quantitative Linguistics* 1.3, pp. 178–188.

Veldhuis, Niek (2011). "Levels of literacy". In: *The Oxford handbook of cuneiform culture*, pp. 68–89.

Veldhuis, Niek C. (2012). "Cuneiform: Changes and developments". In: *The shape of script. How and why writing systems change*. School of Advanced Research Press, pp. 3–23.

Venezky, Richard L. (1970). "Principles for the design of practical writing systems". In: *Anthropological linguistics* 12.7, pp. 256–270.

———— (2004). "In search of the perfect orthography". In: *Written Language & Literacy* 7.2, pp. 139–163.

Voegelin, Charles F. and Florence M. Voegelin (1961). "Typological classification of systems with included, excluded and self-sufficient alphabets". In: *Anthropological Linguistics*, pp. 55–96.

Wang, Yamei and Géraldine Walther (2023). "Mandarin classifier systems optimize to accommodate communicative pressures". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12664–12674.

Wasow, Thomas (2015). "Ambiguity Avoidance is Overrated". In: *Ambiguity: Language and Communication*. Ed. by Susanne Winkler. De Gruyter, pp. 29–47.

Wedel, Andrew, Scott Jackson, and Abby Kaplan (July 2013). "Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change". In: *Language and speech* 56.3, pp. 395–417.

Wedel, Andrew, Abby Kaplan, and Scott Jackson (May 2013). "High functional load inhibits phonological contrast loss: A corpus study". In: *Cognition* 128.2, pp. 179–186.

Weingarten, Rüdiger (2011). "Comparative graphematics". In: *Written Language & Literacy* 14.1, pp. 12–38.

Winter, Bodo, Marcus Perlman, and Asifa Majid (2018). "Vision dominates in perceptual language: English sensory vocabulary is optimized for usage". In: *Cognition* 179, pp. 213–220.

Winters, James and Olivier Morin (2019). "From context to code: Information transfer constrains the emergence of graphic codes". In: *Cognitive Science* 43.3, e12722.

Woods, Christopher (2006). "Bilingualism, scribal learning, and the death of Sumerian". In: *Margins of Writing, Origins of Cultures*. Ed. by Seth L. Sanders. The Oriental Institute of the University of Chicago Chicago, pp. 91–120.

———— (2010). "Visible language: the earliest writing systems". In: *Visible language: Inventions of Writing in the Ancient Middle East and Beyond*. The Oriental Institute of the University of Chicago, pp. 15–25.

Xu, Yang, Khang Duong, et al. (2020). "Conceptual relations predict colexification across languages". In: *Cognition* 201, p. 104280.

Xu, Yang, Terry Regier, and Barbara C. Malt (2016). "Historical semantic chaining and efficient communication: The case of container names". In: *Cognitive science* 40.8, pp. 2081–2094.

Yin, Sora Heng and James White (2018). "Neutralization and homophony avoidance in phonological learning". In: *Cognition* 179, pp. 89–101.

Zaslavsky, Noga, Karee Garvin, et al. (2022). "The evolution of color naming reflects pressure for efficiency: Evidence from the recent past". In: *Journal of Language Evolution* 7.2, pp. 184–199.

Zaslavsky, Noga, Mora Maldonado, and Jennifer Culbertson (2021). "Let's talk (efficiently) about us: Person systems achieve near-optimal compression". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 43. 43.

Zaslavsky, Noga et al. (2018). "Efficient compression in color naming and its evolution". In: *Proceedings of the National Academy of Sciences* 115.31, pp. 7937–7942.

Zaslavsky, Noga et al. (2019). "Semantic categories of artifacts and animals reflect efficient coding". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 41.
Zehentner, Eva (2022). "Ambiguity avoidance as a factor in the rise of the English dative alternation". In: *Cognitive Linguistics* 33.1, pp. 3–33.
Zipf, George Kingsley (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Addison-Wesley.