The Taxonomy of Writing Systems: How to Measure How Logographic a System Is

Richard Sproat
Search
Google, Japan
rws@google.com

Alexander Gutkin
Research & Machine Intelligence
Google, UK
agutkin@google.com

Taxonomies of writing systems since Gelb (1952) have classified systems based on what the written symbols represent: if they represent words or morphemes, they are logographic*; if syllables,* syllabic*; if segments,* alphabetic*; and so forth. Sproat (2000) and Rogers (2005) broke with tradition by splitting the logographic and* phonographic *aspects into two dimensions, with logography being graded rather than a categorical distinction. A system could be syllabic, and highly logographic; or alphabetic, and mostly non-logographic. This accords better with how writing systems actually work, but neither author proposed a method for measuring logography.*

In this article we propose a novel measure of the degree of logography that uses an attention-based sequence-to-sequence model trained to predict the spelling *of a token from its pronunciation in context. In an ideal phonographic system, the model should need to attend to only the current token in order to compute how to spell it, and this would show in the attention matrix activations. In contrast, with a logographic system, where a given pronunciation might correspond to several different spellings, the model would need to attend to a broader context. The ratio of the activation outside the token and the total activation forms the basis of our measure. We compare this with a simple lexical measure, and an entropic measure, as well as several other neural models, and argue that on balance our attention-based measure accords best with intuition about how logographic various systems are.*

Our work provides the first quantifiable measure of the notion of logography that accords with linguistic intuition and, we argue, provides better insight into what this notion means.

1. Introduction

Some time during the third millennium BCE, in Mesopotamia, people discovered, over the course of a few hundred years (Woods, Teeter, and Emberling 2010), that spoken

Submission received: 21 September 2020; revised version received: 21 April 2021; accepted for publication: 3 June 2021.

https://doi.org/10.1162/COLI_a_00409

words and phrases could be represented by marks on a more or less permanent surface. While these marks started out as **logographic** representations of words whose denotations could be easily depicted with drawings, the scribes who developed the first writing soon discovered the rebus principle, and started to use symbols not for what they depicted, but rather for how the symbol was pronounced. This in turn allowed for a much more flexible system, since while it was not always easy to come up with a drawing that evoked the word in question, once the **phonographic principle** was discovered it was much easier to come up with a representation that reflected its pronunciation. Writing systems that (as far as we know) were independently developed in other places and at other times—in Egypt, China, and Meso-America—all followed the same course, and all ended up as a mix of *logographic* and *phonographic* components.

In the oldest writing systems the phonological units represented were usually syllables, though in the case of Egyptian they represented sequences of consonants, ignoring intervening vowels. But over the course of history various phonological units have been represented in different writing systems: syllables; moraic units; consonants; consonants with an inherent vowel but with marks to represent other vowels; or all segments. The typology of phonographic units in writing systems, based on what the systems represent, is thus reasonably clear:

- Syllabic systems. Examples: Modern Yi (Shi 1996), Chinese—as far as the *phonological* unit represented by the individual characters is concerned (Mair 1996).
- Moraic systems. Examples: Japanese Kana (Smith 1996).
- Consonantal systems (abjads). Examples: Semitic scripts, in particular early Semitic scripts (Naveh 1982), such as the early Sinaitic scripts (O'Connor 1996).
- Consonants with inherent vowels, with diacritics to indicate other vowels (abugidas/alphasyllabaries). Examples: Brahmic scripts (Salomon 1996), Ethiopic (Haile 1996).
- All segments (alphabets). Examples: Greek (Threatte 1996), Hangeul (King 1996).

Insofar as these divisions are reasonably straightforward, it is no surprise then that taxonomies of writing systems have typically been based to a large extent on phonographic principles. Gelb (1952), for example, viewed the evolution of writing teleologically, with the earliest systems being logographic, but with syllabaries, consonantal systems (which Gelb viewed as degenerate syllabaries), and finally the alphabet ensuing. Gelb's taxonomy (see his Figure 95, page 191), therefore, is effectively linear. Other authors, such as Sampson (1985, 2012) (e.g., Sampson [1985], Figure 3, page 32) or DeFrancis (1989) have taken less teleological views, preferring instead to present an arboreal taxonomy where, say, segmental systems and syllabic systems are merely on different branches of the tree. Figure 1, for example, is DeFrancis's taxonomy, which again is primarily based on the type of phonological information encoded.

But as we see in Figure 1, DeFrancis's taxonomy also makes another distinction: under the main preterminal node (syllabic, consonantal, alphabetic) we see a further division into whether or not the system is purely phonological, or includes some morphological, or in other words logographic, component. But this replication of the notion of logography on each branch of the tree suggests that phonographic and logographic

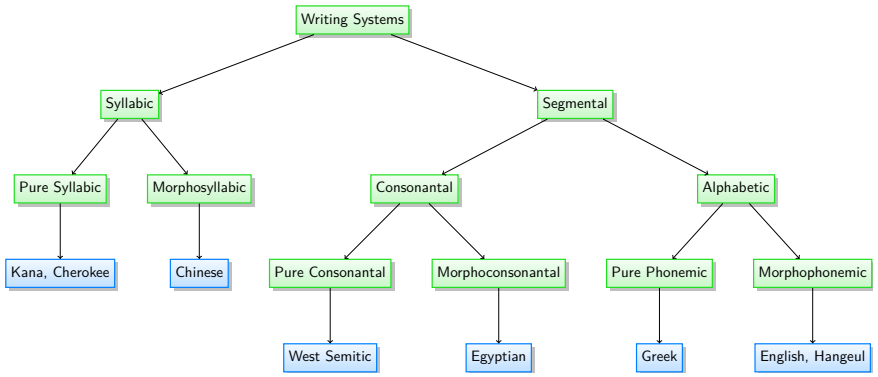


Figure 1
DeFrancis' (1989) taxonomy of writing systems; simplified from his Figure 10, page 58, to focus on what he considers to be true writing systems.

aspects of writing systems are really separate, largely independent dimensions. In principle, a writing system could represent one or another sort of phonological information, and at the same time be more or less logographic. And this in turn suggests that a more apt taxonomy would be planar, with the two dimensions of logography and phonography being on different axes.

Sproat (2000) proposed just such a planar representation, and this system was further developed by Rogers (2005). We present Rogers' revised system in Figure 2, where note that Rogers, like DeFrancis, prefers the term "morphography" to "logography."

Again, the phonographic divisions are reasonably well differentiated on the basis of the kinds of phonological units represented by the writing system, but logography (morphography) seems in principle to be a matter of degree. Chinese, for example, has a fairly heavily logographic system, whereas a highly phonologically simple system like that of Finnish has little or no logography. It thus appears that one could order systems on the basis of how much logographic information they contain.

However, while it seems that this should indeed be possible, neither Sproat nor Rogers provided a way of quantifying how much logography a system contained. The first proposal for a way of quantifying this was made by Penn and Choma (2006), who propose a method that uses sample correlation coefficients to examine the cooccurrence of characters of a writing system in a corpus. We turn in Section 3 to a description of their system. Unfortunately, as we shall also see, their proposal does not work. We then discuss, in Sections 4 and 5.1, an alternative approach based on an attention-based neural model. Section 5.2 presents a simple alternative based on counting the number of cases where there is an ambiguity in how to spell two homophonous words in a dictionary, and Section 5.3 presents information-theoretic measures based on n-gram entropy. In Section 6 we present experiments using our three types of measures on a variety of languages. We also make our code available for others to experiment with.[1]

Note that we focus here almost exclusively on modern languages for which it is possible to obtain a reasonable amount of training data. This means that unfortunately many of the interesting ancient writing systems such as Egyptian, Sumerian, Akkadian, Hittite, or Mayan are outside the scope of the present investigation.

1 https://github.com/google-research/google-research/tree/master/homophonous_logography.

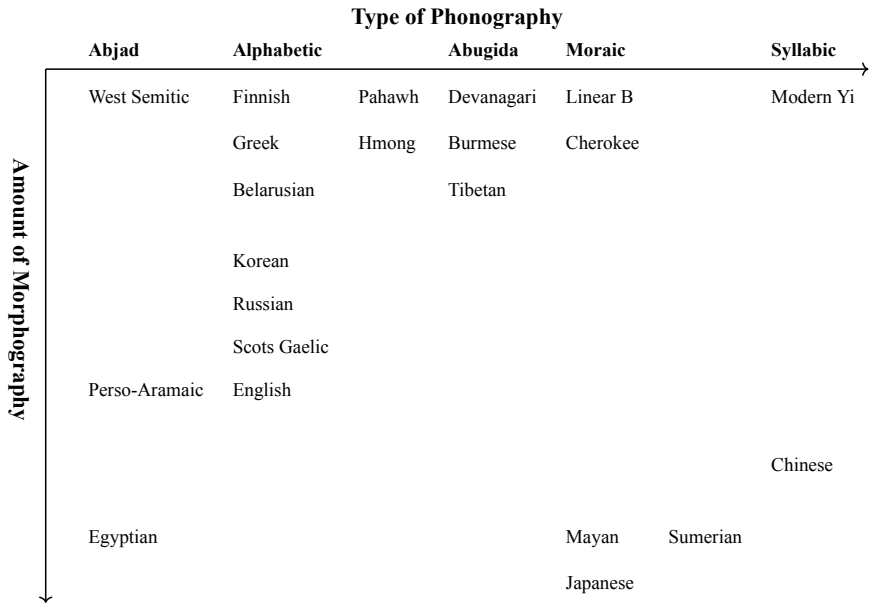


Figure 2
Rogers' (2005) planar taxonomy (his Figure 14.5, page 275), developed based on an earlier proposal in Sproat (2000) (= Sproat's Figure 4.5, page 142).

Before we turn to Penn and Choma's work, however, it is best to be clear about terminology. In Section 2.1, we will review notions of logography as they have appeared in the literature on writing systems, and in Section 2.2 we will provide some operational definitions of what the term means; note that in the immediate discussion here, in giving concrete examples, we will anticipate our **distinct homophones** notion of logography (Section 2.2.1). We will use the term *logographic* throughout this article, but we have noted that DeFrancis and Rogers both prefer a term that evokes the morpheme rather than the word—*morphographic* in Rogers' case—as does Joyce (2011).

As we will see, what we take as the target unit has a large effect on how logographic a system will appear to be. For Chinese, if we view the target as the morpheme, and let us take for example the (Mandarin) syllable *mǎ*, then there are quite a few possible individual characters that could be intended: 馬 ('horse') or 螞 (prefix used with some names of invertebrates) are just two possibilities. If on the other hand we view the target as the word—for example, *mǎyǐ*—then there are far fewer possibilities, and indeed only one is salient: 螞蟻 'ant'. In the former case, where we pick the morpheme as the target, one certainly needs to look outside the morpheme in order to decide how to spell the word. In the case where we pick the word, in the case of 螞蟻 one can decide largely independently of context how to spell it. At least in view of the measure of logography we develop below, it makes a difference what unit we intend. In the ensuing discussion we will use the terms *morpheme* and *word* as appropriate, but consistently use the terms *logography* and *logographic* in either case. Also, in order to avoid awkward locutions like *word or morpheme*, we will use the term *word* when nothing important hinges on the difference. Finally, while the planar taxonomy just discussed clearly suggests that logography is a matter of degree, we will for ease of locution use phrases like *purely*

phonographic to denote writing systems that have a very low degree of logography, and are therefore situated near the top of the plane in Figure 2.

2. What Defines "Logography"?

Given how prominently logography figures in the study of writing systems, one might expect that the notion has been clearly defined. Unfortunately this is rarely the case. In this section we start out by reviewing the definitions, such as they are, that have been given by various scholars over the past seventy years. We then attempt to tease apart the various notions that have been hinted at in the literature, and discuss what one would need to do to operationalize these notions. As we shall see, we will concentrate in this article on one notion that we will term the **distinct homophones** notion of logography.

2.1 Prior Definitions in the Literature

How do writers on writing systems define the notion "logographic"? The quotations in Table 1, derived from a reasonably comprehensive set of works starting with Gelb (1952), illustrate the kinds of definitions that have been offered. Few of the passages quoted in Table 1 really count as definitions. The most one generally finds is that authors point to an example or two that fits their notion, and hope that the reader's common sense will allow them to know an instance when they see it. Even the more extensive description by Pope (1975) begs the question of how one determines that something *is* a logogram representing the word itself, as opposed to just a way to write a particular sequence of phonemes that happens to be the pronunciation of a particular word. While he cites the specific case of Chinese, the majority of Chinese characters do not really constitute "a sign for a complete word" since Chinese characters themselves are usually decomposable into a part that indicates something about the meaning, and a part that indicates something about the pronunciation. So since most Chinese characters are not single signs, it is not clear how one would distinguish this case from the case of a word spelled in a purely phonographic script where the "word sign" in this case is just the sequence of phonograms used to spell the word.[2]

The best description and the closest to a formal definition is the last one given, that of Handel (2019). This is not surprising given that the entire theme of his book is logography and how logographic writing systems have been adapted. Even then Handel's definition begs the question what he means by "basic graphic elements" since the logographic Chinese characters that his work focuses on are certainly not "basic graphic elements" in and of themselves, being as they are mostly constructed from more basic elements, where those elements are nearly always, if anything, either phonographic or *semasiographic*, not logographic. Interestingly, Handel's functional definition of logography closely accords with the **distinct homophones** notion that we develop most thoroughly in this article. But as we argue below, this is only one possible way of viewing logography that is consistent with the way the term has been used in the literature.

2 Note that this is precisely the reason that DeFrancis (1989) argues that Chinese writing is not really logographic. Of course not all Chinese characters are thus decomposable: 人 *rén* 'person' is not. And in Japanese *kunyomi*, where a Chinese character is pronounced as a native word such as 神 *kami* 'god', any such semantic-phonetic decomposition in the Chinese character is useless for determining the pronunciation of the Japanese word. Perhaps Pope could have defined his notion precisely by focusing on such cases, but he did not do that.

Table 1
Prior "definitions" of *logography* given in the literature.

Gelb (1952, p. 65): "The signs used in the earliest Uruk writing are clearly word signs limited to the expression of numerals, objects, and personal names. This is the stage of writing that we call logography or word writing and that should be sharply distinguished from the so-called 'ideography."' Further (p. 99): "Logograms, that is signs for words of the language."

Moorehouse (1953, p. 26, fn. 9): "A sign so used may be called a logogram: it is a sign attached to a particular word, though without reference to its meaning (which would make it an ideogram) or its sound (a phonogram)." This is in his discussion of the Hittite use of Akkadian ⟨ABU⟩ 'father' for Hittite *attash* 'father'.

Diringer (1958) does not use the term "logograph," but rather the now disfavored term "ideograph," but he hints (p. 43) that this is not really an appropriate usage: "At a second stage, the symbols represented also ideas; signs were borrowed from those denoting words related in meaning, for example the solar disk came to represent also the ideas of 'day' and 'time.' Characters used in this way are called, though not quite correctly, *ideographs*; they were, to be more exact, word-signs. . . "

Pope (1975) gives something closer to an operational definition of logography when he writes (p. 203): "*logogram* a sign for a complete word, differing from a *determinative* in that it furnishes additional information instead of classifying information already given. Chinese characters are logograms, and Chinese can be called a logographic script. But most, perhaps all, other scripts contain a class of logograms. English examples include £, $, =, + as well as all the numeral signs. Abbreviations, though composed of *phonograms*, are logographic in function."

Sampson (1985, p. 33): "logographic systems are those based on meaningful units"

Coulmas (1989) does not really define the term, but implies that a logogram is a sign used to represent the word or meaning (see e.g., his discussion on page 78).

DeFrancis (1989) defines the term much as others do, but rejects it as inappropriate, particularly for Chinese, since for him scripts always have strong phonetic components. He characterizes Chinese as *morphosyllabic*, meaning that each character denotes a syllable as well as some aspect of the meaning of the morpheme.

Drucker (1995, p. 14): "Writing systems . . . may also be logographic, in which case the written sign represents a single word."

Harris (1995) reviews a number of typologies, but never provides a definition of what the term "logographic" denotes.

Daniels (1996b, p. 9): "Istrin's[a] 'ideograms' do not in fact record 'ideas' (Gelb rightly banished the term from our science, preferring *logogram*) but rather individual words or their significant parts."

Sproat (2000) largely followed in DeFrancis's footsteps in rejecting the categorical distinction between logographic and phonographic scripts.

Coulmas (2003, p. 47): "Being logograms, the signs refer to these words in their entirety, that is, the graphic complexity of the signs is not related to the internal structure of the words."

Rogers (2005, p. 14): "When we get to Chinese in chapter 3, we will meet a writing system where the primary relationship of graphemes is to morphemes. Such a system can be called **morphographic**, and those graphemes can be termed **morphograms**" (boldface original).

Robinson (2007, p. 13): "Europeans and Americans of ordinary literacy must recognize and write about 52 alphabetic signs, and sundry other symbols, such as numerals, punctuation marks and whole-word semantic symbols, for example +, &, £, $, 2, which are sometimes called logograms."

Dehaene (2009), citing Frith (1985), takes a neurological/psychological view of the term and uses it to denote the phase of learning to read (really a stage prior to learning to read) where children recognize words in terms of their overall shapes. The classic instance of this is children who can "read" the brand name *Coca Cola* as written in its traditional cursive form. In common with the grammatological definitions of the term is the notion that the written form represents a whole word.

Daniels (2018, p. 155): The closest thing to a definition is here: "logogram: a symbol (often a pictogram) denoting the meaning but not the pronunciation of a word or morpheme"

Handel (2019, pp. 7–8): "In a logographic system, the basic graphic elements represent meaningful elements of the spoken language, so that identically pronounced but semantically contrastive elements have distinct graphic representations."

In the next section we explore some phenomena that would seem to fall under the various notions of logography that have been hinted at in the literature, and thereby try to put the notion on a more rigorous and operationalizable footing.

[a] Viktor Aleksandrovich Istrin (1906–1967), Soviet philologist. See e.g., Istrin (1965).

2.2 Specific Notions of Logography

As we have just seen in the previous section, nobody really *defines* what the term *logography* (or *morphography*) means. Having said this, it is clear that a variety of phenomena are intended. This section tries to develop a taxonomy of such phenomena.

2.2.1 Different Words Should Be Spelled Differently. The most obvious idea lurking behind the notion of logography is the idea that words that are different *even if they sound the same* should be spelled differently.[3] It is, for example, the basis for Sampson's (1985) idea that English orthography is at least partly logographic. The basis for the spelling differences can be various. They may be due to historical sound change rendering two words with originally distinct pronunciations homophonous. Or they may have been artificially introduced, sometimes due to (pseudo)etymological considerations. Let us term this the **distinct homophones** notion of logography. As noted previously, this is also the notion that the recent work of Handel (2019) promotes.

Japanese provides many instances of this notion of logography.[4] Thus Japanese has many words that are homophonous, but are written using different kanji. To take a simple example, 結婚 'marriage' and 血痕 'bloodstain' are both pronounced *kekkon*, identical even in pitch accent in Tokyo Japanese (both are unaccented). More subtle are cases where what seems to be the same word, in that it has the same pronunciation, and the same general meaning, nonetheless has more than one kanji representation, where the different spellings seem to relate to different senses of the word. For example, the normal word for 'town' in Japanese is *machi*, typically written 町. This was derived from a Chinese character meaning a ridge between fields (田) and came to be used in Japan for an administrative area. However another spelling is also used, 街, which in Chinese means 'street'. The former is associated with the meaning of 'town' as an (administrative) area. Thus if you are talking about where someone lives, you would use 町. The other spelling is more associated with the sense of a town as a collection of buildings along a street: when talking about shopping in town, one could use 街. As another example, compare 匂い *nioi* and 臭い *nioi*. Both mean 'smell', but the former is more associated with 'fragrance', whereas the latter has the connotation of a bad smell, deriving as it does from the Chinese character 臭 *chòu* 'stink'. Many further such cases can be found in Halpern (2013). In principle one might simply consider these to be different but homophonous words, and it is hard to distinguish them from 結婚/

3 Some readers may object to our use of the term *spelling* here since for them it may have the connotation of spelling in an alphabetic writing system. A more "proper" term might be *graphematic representation*. However, this is a somewhat awkward locution so we will use the term *spelling* throughout as a proxy for this with the meaning of 'how one writes a given word or morpheme in the orthography of the language'.

4 We focus here on cases involving kanji (Chinese character) spellings, though Japanese also presents a number of cases that could be termed logographic that involve kanji, hiragana, and/or katakana spellings.

See the introductory chapter of Sansom (1928), and also Handel (2019), for a lucid discussion of how Chinese writing was adapted to Japanese. Sansom discusses how a given kanji can represent multiple different Japanese words as well as the original Chinese word, and how the same Japanese word may be represented by more than one kanji. He also notes, as we do below, that Japanese writing in some ways approaches "ideography."

血痕 other than the vague sense that these are still the same word, where the spelling difference focuses on different nuances of that word. One finds somewhat similar cases in English: For example, though they would probably count as different words now, *brake* is etymologically derived from *break*—that is, a device for breaking the movement of a vehicle.

What is crucial here is the notion that different words (or senses of the same word) should have different spellings, even if they are pronounced the same. This is more precise than the vague notion presented in the literature that a given word is written, at least in part, with a symbol that somehow represents that whole word. For unless one has another word with the same sound, but a different written form, how would one know that the written form of the word is not simply an idiosyncratic way to write the *pronunciation* of the word? If I have a written form ⟨  ⟩ used to represent the word *pig*, how does one know this is not just a way of writing the phoneme sequence or syllable /pig/? Presumably one could only be sure of this by showing that in general distinct words that sound the same, nonetheless end up having different spellings, by virtue of the fact that the words are, after all, different. Thus in the case at hand, the word *pig* that is synonymous with *ingot* (as in *pig iron*), should in that case be written with a different symbol, say ⟨⟩.[5,6]

The **distinct homophones** notion of logography is reasonably straightforward to operationalize: One merely needs to compute to what extent a writing system spells words that are pronounced the same in distinct ways, since this is a clear indication that spelling is being used to distinguish among words. It is this notion of logography that the measures proposed in this article most clearly address.

2.2.2 The Same Morpheme Should Be Spelled the Same. The flip side of cases just described in Section 2.2.1 involves the idea that the same morpheme should be spelled the same in all contexts, no matter the actual pronunciation in any given instance. Note that here the notion of *morpheme* is more accurate since what we are talking about here are cases where a morphological component should be spelled the same even when morphophonological changes obscure the phonological form. Let us term this the **uniform spelling** notion of logography.

5 While this example may seem arcane, this is in fact a real issue in Egyptian, where it is often hard to tell whether one should consider a triliteral root as a logograph or as merely phonographic for that sequence of three consonants. Thus the scarab beetle ⟨  ⟩ is presumably logographic when it writes the word for scarab, but phonographic when it is used for *ḫpr* 'become'.

6 Also, while this may seem all rather obvious, it is far from trivial, since even with writing systems that clearly involve logography according to the above procedure, there is evidence that fluent readers treat logograms as if they were phonographic. A hint along these lines is that in proofreading it is often very difficult to catch errors involving obviously distinct words, such as *there* and *their*, that happen to be pronounced the same. So consider our Japanese examples from above, 結婚 'marriage' and 血痕 'blood stain', both pronounced *kekkon*. Sproat (2000) in Section 5.2.2 discussed work by Horodeck (1987) and Matsunaga (1994) that showed that Japanese readers were significantly less likely to catch errors in text that involved homophonous words with different spellings such as the 結婚/血痕 example just given, than in cases where the incorrect word had a different pronunciation from the correct word. What this suggests is that fluent readers develop strategies whereby written symbols, even ostensibly logographic ones, map directly to their pronunciations, effectively bypassing the mapping between the written form and the word it denotes. Thus readers will miss cases where the word is misspelled, but the resulting pronunciation is the same as that of the correctly spelled target word. Thus, though it is clearly the case that 結婚 and 血痕 are not simply spellings of the phonemic form *kekkon*, nonetheless fluent Japanese readers effectively treat them that way. In a similar vein, readers of English will be familiar with the fact that it is often hard to catch spelling errors that involve homophones: confusions of *to* and *too*, or *there* and *their*, are often hard to catch.

It is this notion of logography that Chomsky and Halle (1968) were implicitly appealing to when they claimed that English orthography is a "near perfect" representation for English, focusing on cases like *telegraph* /ˈtɛləgɹæf/, *telegraphy* /təˈlɛgɹəf-/,[7] and *telegraphic* /tɛləˈgɹæf-/, where the spelling remains constant in the various derived forms even though the phonetic form of the stem changes.

This is also the sense in which (the phonographic elements of) Egyptian and Semitic scripts can be considered logographic: In their purest form, the underlying (root) morpheme is spelled with the same set of consonants no matter what vowels are used, so that ⟨KTB⟩ representing the root meaning 'write' would be used, thus spelled in a variety of derived words with different phonetic forms. Egyptian phonetic spelling and the earliest Semitic scripts were fairly pure in their keeping of this convention. The purity was lost to some extent by the introduction of *matres lectionis*, the use of consonant symbols to represent (long) vowels: These necessarily had the effect of breaking up the consonants of the root, thus somewhat obscuring the same root in phonologically distinct forms. For example, Hebrew כתב ⟨ktb⟩ /katav/ 'he wrote', contrasts with יכתוב ⟨yktwb⟩ /yaxtov/ 'he will write', where besides the prefix ⟨y⟩, a vav ⟨w⟩ intercedes between the ⟨t⟩ and ⟨b⟩ of the root. The later development of systems of *pointing*, such as the Masoretic *nikud* for Hebrew (Ravid 2005), on the other hand, did not have this effect, since the points were always written as additional diacritics around the consonants, with the explicit intent of *not* breaking up the consonant sequence.

Similarly, it is the sense in which Middle Persian languages used **heterograms** (also called **aramaeograms**)—Persian words written using Aramaic root spellings, so that the word for 'king', Persian /ʃax/ would be *written* using the Semitic word ⟨MLK?⟩ (Skjaervo 1996), since /ʃax/ and ⟨MLK?⟩ both *mean* the same thing, and the Semitic spelling provided a constant spelling for the Persian morpheme. The general phenomenon of writing a word in one language but with the intention that it be read in a different language is termed *alloglottography* (Rubio 2006; Kudrinski and Yakubovich 2016). According to Gershevitch (1979), the gradual switch to the phonographic Imperial Aramaic-derived spelling system from logographic Elamite cuneiform occurred at some point during the fifth century BCE. Some aramaeograms even occur with Aramaic possessive suffixes despite representing a Persian noun without a possessive, for example, Aramaic ⟨?MY⟩ for *my mother* standing for /māt/, or ⟨BRH⟩ for *his son* representing Persian /pus/ for *son* (Rubio 2006).

Japanese also provides some interesting examples that seem to be related to the **uniform spelling** notion of logography. Most similar to Middle Persian heterograms are the numerous instances in which the same kanji is used to represent either a native Japanese word, or a Chinese borrowing. Since a majority of kanji have one or more Chinese-derived pronunciations, and one or more native pronunciations, this case is very common indeed. For example, the spelling we have encountered already, 町 *machi* 'town', can also be used to write the Chinese-derived morpheme *chō*, also meaning 'town'. Given that the character 町 is, not surprisingly, common in place names, one of the difficulties in Japan is knowing whether in a particular place name this character should be pronounced as *machi* or *chō*. The other character for 'town' that we encountered, 街 *machi*, can also be used to represent the Chinese-derived word *gai*, 'town, street'.

These cases clearly involve the same spelling being used for two morphemes—one native, one Chinese-derived—which are nonetheless related in meaning. This carries the

7 In the first author's pronunciation.

notion of **uniform spelling** to a new level in that we are now dealing with a case where the *same meaning* is spelled in the same way: aspects of Japanese writing thus border on *semasiographic* rather than merely *logographic*. But Japanese carries this even further to cases where the same spelling is used to represent two distinct words that merely have a vague semantic connection. For example the verb meaning to 'get off, disembark, alight', written 降ります *orimasu* is spelled the same way as the verb for precipitation (rain, snow) 降ります *furimasu*. These are clearly different words, but share the sense of something descending from something else.

While **uniform spelling** seems to be a valid characterization of what is intended by some discussions of logography, it is also much harder to operationalize than the **distinct homophones** notion. In the latter case, one merely has to have a way of computing the pronunciation of the differently spelled words, taking at face value the assumption that words that are *spelled* differently are indeed intended to be distinct. For **uniform spelling** one needs to have access to an underlying notion that two tokens that are spelled the same are nonetheless both different phonologically, yet at the same time ultimately the same morpheme. This can be difficult to compute depending on the language and resources. Thus while Japanese 町 *machi*/*chō* 'town' clearly represent different morphemes, nonetheless they represent the same meaning, so at some level they represent a consistent spelling of the same entity, which happens to be pronounced in different ways in different contexts. The difficulty is that it is hard to predict the context in which one would get one pronunciation or the other. For example in town names ending in 町 one often just has to know how the name is pronounced in any given instance. Databases of such toponyms exist, but nothing that is open source. To be sure, information relevant for the **distinct homophones** notion is easier to derive for some languages than others. Thus for Semitic languages, if we know the consonantal root from the spelling, we can (usually) be reasonably sure that it is intended to represent the same morpheme, and at the same time if we have the diacriticized text, we can know whether or not the pronunciation differs in two different cases. Thus, we could compute values for the degree of **uniform spelling** homography for Hebrew (see below). However developing materials is sufficiently complicated for a range of languages that we have decided in this article to forgo this definition and concentrate on the **distinct homophones** definition. We therefore leave **uniform spelling** for future work.

We now turn in the next section to a discussion of the only work to date that has attempted to quantify the degree of logography in writing systems, namely, that of Penn and Choma (2006). We will show that their results are not replicable and seem to depend on an artifact relating to different corpus sizes used for the two languages they compared.

3. Penn and Choma's Correlation Coefficient-based Measure

The proposal by Penn and Choma (2006) is based on the intuition that logographic symbols, insofar as they are associated with particular words, are also therefore associated with the word's meaning. This in turn leads to the expectation that their distribution in text should be "clumpy," since semantically connected words tend to co-occur. Thus for the Chinese character 牛 (*niú* 'cow') one would not be surprised to find 草 (*cǎo* 'grass') nearby, whereas 市 (*shì* 'city') would be less expected. On the other hand, in a purely phonographic system, the symbols are not intrinsically associated with meaning, and thus one has no expectation of such an association: The distribution should be less "clumpy" and symbols should be broadly correlated with one another.

Before proceeding, it is worth asking which of our notions of logography—**distinct homophones** or **uniform spelling**—Penn and Choma's measure was intended to address. Penn and Choma do not present any formal definition of what they mean by logography, but they clearly intend that in a highly logographic system the symbols should convey semantic information, and thus be expected to show clumpiness; whereas in a system with very little logography, the symbols have no inherent semantics. While clearly not identical to it, this is at least consistent with the **distinct homophones** notion, the main target of our study. In a purely phonographic system, words that sound the same will be spelled the same no matter their semantic relationship, since the symbols do not relate to meaning. Thus the meaning differences would be neutralized in the spelled form, leading to the expectation that the spelling would not be associated strongly with any particular topic. In contrast, in a logographic system the words in question should be spelled differently, due to the meaning differences, and one would expect a stronger association to particular topic-related semantic clumps.

The clumpiness noted above, Penn and Choma propose to capture using sample correlation coefficients between two random variables X and Y

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma(X)\sigma(Y)} \tag{1}$$

defined as the sample covariance $\mathrm{cov}(X, Y)$ of two random variables divided by the product of their standard deviations σ, where

$$\mathrm{cov}(X, Y) = \frac{1}{n-1} \sum_{0 \le i,j \le n} (x_i - \mu_i(X)) \left(y_j - \mu_j(Y)\right) \tag{2}$$

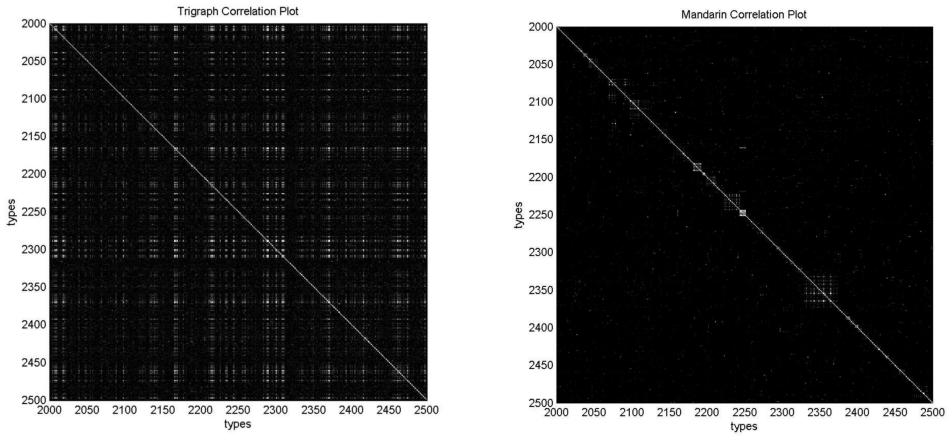
and

$$\sigma(X) = \sqrt{\frac{1}{n-1} \sum_{0 \le i \le n} (x_i - \mu_i(X))^2} \tag{3}$$

where μ is the mean of the variable. Per their description (page 119), "each grapheme type is treated as a variable, and each document represents an observation."

A reasonable comparison of two writing systems would involve a system with a large degree of logography, and one that has little or no logography, but the phonographic symbols constitute a set of the same order of magnitude as the logographic system. So for example, one could pick Chinese, with a *morphosyllabic* system (DeFrancis 1989), where each character represents a syllable, but also gives some indication of which particular morpheme is involved; and Modern Yi, a pure syllabary consisting of a few hundred syllabic symbols that is used to write the Yi language (Tibeto-Burman, Southwestern China) (Shi 1996). While Chinese corpora are easy to come by, unfortunately Yi corpora are not, and so rather than use Yi, Penn and Choma approximated a purely phonographic system by what they term *trigram English*. Take an English text, and for each word, divide it into sequences of three letters, leaving any trailing letters in their own group. Thus *grapheme* would be divided into *gra*, *phe*, and *me*.

For text corpora, they used an unspecified Chinese news corpus, and for (trigram) English the Brown corpus (Kučera and Francis 1967). The correlation coefficients can be depicted as heatmaps where each axis represents characters, and each position i, j in the map has a brightness proportional to the coefficient of the ith and jth characters:



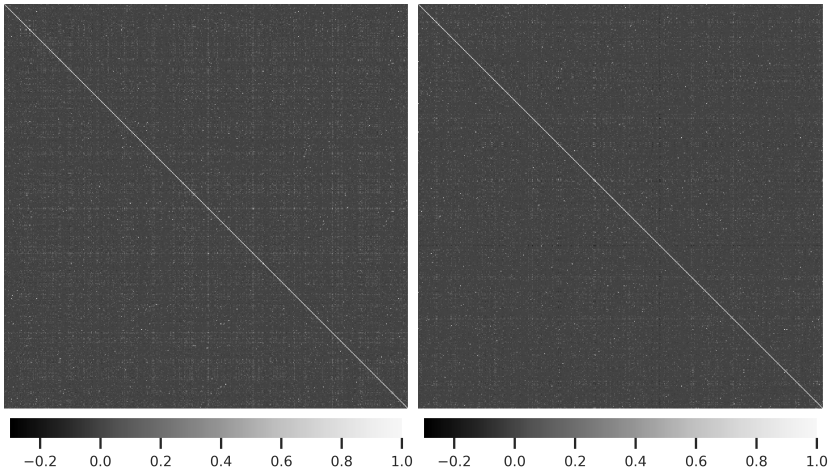
(a) Trigram English (b) Chinese

Figure 3
Penn and Choma's results for trigram English and Chinese (Mandarin), their figures 3 and 4, respectively.

Brighter dots are associated with more strongly correlated characters. Obviously in such a diagram, the diagonal is always bright since all characters correlate with themselves. Penn and Choma's results for trigram English and Chinese are shown in Figure 3, where they note that the trigram English plot is considerably "brighter" than the corresponding Mandarin plot. This suggests that the phonographic system with a low degree of logography (or in their case, a simulation of such a system) has more promiscuous correlations between symbols than is the case for Chinese, which has a much less bright plot, and where the points of brightness seem to be clumped more locally. This is in accord with the intuition sketched at the beginning of this section. More promiscuous correlations imply higher numbers, and as Penn and Choma state (page 120):

> By adding the absolute values of the correlations over these matrices (normalized for number of graphemes), we obtain a measure of the extent of the correlation. Pervasive semantic clumping, which would be indicative of a high degree of logography, corresponds to a small extent of correlation—in other words the correlation is pinpointed at semantically related logograms, rather than smeared over semantically orthogonal phonograms. In our example, these sums were repeated for several 2500-type samples from among the approximately 35,000 types in the trigram English data, and the approximately 4,500 types in the Mandarin data. The average sum for trigram English was 302,750 whereas for Mandarin Chinese it was 98,700. Visually, this difference is apparent in that the trigram English matrix is brighter than the Mandarin one. From this we should conclude that Mandarin Chinese has a higher degree of logography than trigram English.

Unfortunately, as it turns out, Penn and Choma's results are not replicable, and one of the problems comes down to their choice of corpora. The Brown corpus and newswire text are rather different, not the least in the lengths of the documents. The Brown corpus (Kučera and Francis 1967) was designed to consist of a set of documents of about 2,000 words each, and assuming a mean word length of about 5 letters, this works out to about 4,000 three-letter sequences per document. The LDC Chinese Gigaword corpus



(a) Trigram English (b) Chinese

Figure 4
Penn and Choma's method applied to the Chinese and (trigram) English Bibles, showing correlation coefficients for 500 characters.

(Graff and Chen 2005), to take a fairly typical example of a Chinese news corpus, on the other hand, has an average document length of about 450 characters. Thus the trigram English documents have almost 9 times as many characters on average as the Chinese documents. A reasonable hypothesis, then, is that the reason the trigram English plot is "brighter" is simply that there is more chance for any given pair of "characters" to co-occur within a document than is the case in Chinese. If so, their result has nothing to do with what the characters denote, and therefore cannot be used as a measure of the amount of logography in a system.

To confirm this we used the Chinese and (trigram) English texts from the Bible Corpus (Christodoulopoulos and Steedman 2015). We took a chapter as a document, and trigrammed the English portion of the corpus as Penn and Choma did in their experiments. On average there are then 1,100 "letters" per document for English and 780 characters per document for Chinese. Figure 4 shows the plots for trigrammed English and Chinese. It will be seen that there is very little difference in "brightness" between the plots.

If we then take the English Bible and group six consecutive chapters into a "document," this yields about 6,600 "letters" per document, or about 8.5 times the number that is in a Chinese document. In this case the correlation coefficient for trigram English is considerably "brighter"—there are more bright dots in this plot than in the previous two plots—replicating Penn and Choma's result. See Figure 5.

In terms of overall correlations, we computed total absolute correlation of 500 randomly selected characters, averaged over five runs for Chinese, trigram, English, and, as an example of another largely phonographic script, Korean. For all languages we compared two cases: one where a document is a single chapter, and one where it is six consecutive chapters combined, as in the discussion above. Results are shown in Table 2. Also shown in that table is the number of distinct characters, or in the case of trigram English "characters." For Korean, we used as characters the whole *hangeul* syllable, rather than the individual *jamo* letters in order to get a character set size that is

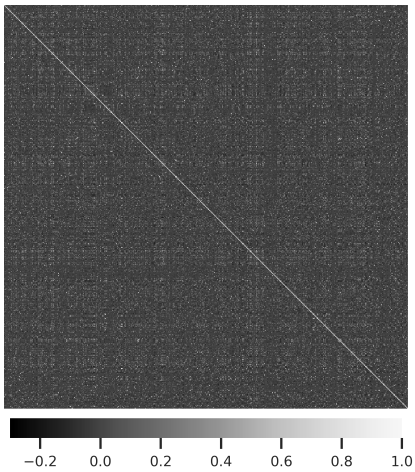


Figure 5
Penn and Choma's method applied to a version of trigram English where six consecutive chapters are combined into a "document"; plot shown for 500 characters. We recommend magnifying the figure to see the difference between this and the previous two plots in Figure 4.

Table 2
Summed absolute correlations averaged over five runs on the Bible corpus for three "scripts": Chinese, trigram English, and Korean syllables. Results are given for two definitions of "document." Also shown are the number of distinct characters in the Bible for each "script."

Language	Doc. = 1 ch.	Doc. = 6 ch.	# distinct chars.
Chinese	5,828	14,859	3,177
3-gram English	6,386	15,116	3,194
Korean	8,508	18,200	1,249

roughly in the ball park of the number of distinct Chinese characters in the corpus. (This differs from how we process Korean for our own experiments discussed later on; see Section 6.1 for more detailed discussion.) The summed correlations for trigram English and Chinese are remarkably similar, once document size is balanced. Korean has overall higher values, but as we see in the final column of Table 2, it also has a smaller character set than the other two. This means that there is more opportunity for two randomly selected characters to co-occur in a document, so one would expect an overall higher correlation value. Penn and Choma's results can be completely explained by document size and character set size.

So Penn and Choma's result turns out to have nothing to do with the function of the symbols in Chinese versus trigram English, but rather is an artifact of differing text sizes. And this is in turn rather unsurprising. Penn and Choma's method is an extrinsic method that purports to rely merely on the distribution of symbols in a corpus. A priori this seems unlikely to be able to discover the function of those symbols, any more than the distribution of symbols is likely to tell you that a given symbol system represents language (i.e., is a true writing system) versus some other kind of information (cf. Sproat 2014).

To be sure, traditional approaches to decipherment have often relied on a very simple extrinsic measure—the size of a symbol set—for making an initial guess as to

what kinds of information the symbols represented. For example, Pope (1999, page 138) describes the early work of A. H. Sayce, one of the pioneers of Luvian hieroglyphic decipherment, who in 1876 suggested that the script must be a syllabary, with an ideographic element present as well, based, among other evidence, on the close similarity between the symbol inventory sizes between Luvian and the recently deciphered Cypriot syllabary. Thus, more generally, a script with only twenty symbols is probably a consonantal system or an alphabet; one with two hundred symbols is probably a syllabary; and one with several hundred or a few thousand symbols is probably some sort of logographic system (Daniels 1996a). But such methods are really only useful for crude initial guesses, and in any case are easily fooled: the Modern Yi syllabary would look like a logographic system, according to such an approach.

So it seems unlikely that statistical methods based solely on the distribution of symbols, no matter how sophisticated, can be very informative about whether a system is logographic or to what degree it is logographic. We need rather to consider not only the symbol but some representation of the linguistic information it encodes. We turn in the next section to a proposal for a measure that takes this into account.

4. Attention-based Classification

In a completely regular phonographic system, such as the Finnish writing system, there is rarely if ever any ambiguity about how to write a word (Aro 2017). Once one knows the pronunciation of the word to be written, the spelling can be derived directly from it. What logography introduces is ambiguity into that process. It is no longer enough to know what the pronunciation of a word is—one must also know which specific word among the several that may share the same pronunciation is intended. In logographic systems, according to the **distinct homophones** notion, spellings are used to distinguish words that are pronounced the same. To take an example from English, in order to know how to write a word pronounced /greɪt/, one needs to know whether what is intended is the word written ⟨great⟩ or the one written ⟨grate⟩. As Sampson (1985) argued, English spelling is at least somewhat logographic precisely in making these kinds of arbitrary distinctions.

Chinese offers another, and perhaps more consistent, method of distinguishing between homophonic words, with the use of so-called semantic radicals. In Mandarin, *pípá* can be one of two words, 琵琶 the Chinese lute, or 枇杷 'loquat', each written with two characters since there are two syllables. Both involve the same phonetic components, 比巴 *biba*, which gives a hint at the pronunciation of the whole word. The only difference is in the semantic radical, repeated on both characters, which is 王王, used to denote musical instruments, in the case of 'lute', and 木 'tree' in the case of 'loquat'.

Choosing the appropriate spelling for homophones such as *great/grate* or 琵琶/枇杷 thus requires knowing which word is intended. Usually this is determinable from the linguistic context in which the homophone occurs, so an operational definition of logography can be simply the extent to which one needs to look at the context of a word in order to determine how to write it. In the Finnish writing system, there should be very little need to look beyond the word itself, whereas for English and Chinese it will frequently be necessary to look in a broader context. This operational definition can be turned into a computational definition by considering what a computational model of the relation between sound and spelling would need to do in order to correctly determine spellings.

One instance of such a model is a neural sequence-to-sequence model with an attention mechanism (Mnih et al. 2014; Bahdanau, Cho, and Bengio 2015). Such a

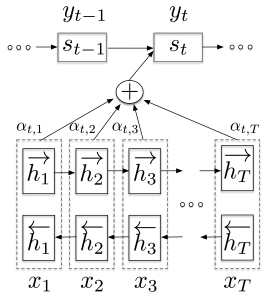


Figure 6
The Bahdanau attention model (Bahdanau, Cho, and Bengio 2015, Figure 1, p. 3): $x_1, x_2, \ldots, x_T$ represent the inputs, the h represent the annotations, and s the hidden states. As described in the text, output y_t is predicted from output y_{t-1}, the previous state s_{t-1} and the sum over the weighted inputs from the annotations.

model is an instance of an encoder-decoder framework. A discrete input sequence—for example, words in a text, or phonemes in a phoneme sequence—is read into the encoder and then embedded as a sequence of vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ in a continuous vector space. The task of the decoder is to predict the next symbol of the output, given the inputs and the previous outputs—that is, to find the probability p of predicting output y_i given the history $y_1, \ldots, y_{i-1}$ and the input vectors $\boldsymbol{x}$: $p(y_i|y_1, \ldots, y_{i-1}, \boldsymbol{x})$. In the model of Bahdanau, Cho, and Bengio (2015), this is modeled as a nonlinear function g, which takes as input the previous output, a context vector c_i, and a hidden state s_i:[8]

$$p(y_i|y_1, \ldots, y_{i-1}, \boldsymbol{x}) = g(y_{i-1}, s_i, c_i) \tag{4}$$

where the hidden state $s_i = f(s_{i-1}, y_{i-1}, c_i)$ uses another nonlinear function f and c_i is in turn defined in terms of a sequence of *annotations* $H = h_1, h_2, \ldots, h_T$,

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j \tag{5}$$

where each annotation is weighted with a probability α_{ij} defined as

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{6}$$

where the energy $e_{ij} = \alpha(s_{i-1}, h_j)$ is an *alignment model*, which provides a score for the match between the inputs around position j and the output at position i. The annotations H are derived from the concatenation of the forward and backward hidden states $\overrightarrow{h_j}$, $\overleftarrow{h_j}$, which allows the system to encode at each position information from the preceding and following words. The model is shown in Figure 6, reproduced from Bahdanau, Cho, and Bengio (2015, Figure 1, page 3).

8 Note that "Bahdanau attention," as it is often called, is only one of several attention mechanisms that have been defined (see for example Xu et al. 2015; Luong, Pham, and Manning 2015; Raffel et al. 2017; Vaswani et al. 2017; Deng et al. 2018; Gülçehre et al. 2019). We have opted for it here because it is simple and intuitive to interpret. More sophisticated attention models are considered in Section 6.7.

Table 3
Opening sentence of the Book of Genesis with phonetic form on the input side and spelling on the output.

Input:	ih0_n dh_ah0 <targ> b_ih0_g_ih1_n_ih0_ng </targ> g_aa1_d k_r_iy0_ey1_t_ah0_d dh_ah0
Output:	beginning

As Bahdanau et al. put it (2015, page 4), "the probability α_{ij} , or its associated energy e_{ij}, reflects the importance of the annotation h_j with respect to the previous hidden state s_{i-1} in deciding the next state s_i and generating y_i. Intuitively, this implements a mechanism of attention in the decoder." In practical terms, attention reflects the importance each portion of the input has for predicting each output symbol.

In this article we use Bahdanau's attention mechanism with the recurrent neural network (RNN) in the encoder and decoder using gated recurrent units (Cho et al. 2014). Unlike the original Bahdanau model shown in Figure 6, our model uses unidirectional, rather than bidirectional (Schuster and Paliwal 1997), RNN in the encoder. The model is trained to learn to spell a word, given its phonetic form and the phonetic form of the sentence context. The training can also be inverted so that the model learns to pronounce a word given its spelling and the spelled words in context. This is familiar in speech technology as the **grapheme-to-phoneme** problem (Milde, Schmidt, and Köhler 2017; Yolchuyeva, Németh, and Gyires-Tóth 2019), and indeed sequence-to-sequence models have come to be widely adopted for various pronunciation and text-normalization problems in speech—compare the RoadRuNNer text normalization system (Zhang et al. 2019) we use in Section 6.5.1. This latter direction is, however, of less interest to us for the present purposes, because our goal is to measure the amount of context that is needed to determine how to *spell* a given *pronunciation*.

The training is set up so that the spelling of a word is presented paired with the phonetic form of the sentence, with special symbols <targ> and </targ> surrounding the target word's phonetic form in the input. For example, consider the opening sentence of Genesis shown in Table 3 using the ARPAbet phonetic transcription from the CMU Pronunciation Dictionary,[9] with underbars linking the phonemes within a word. In practice, the input context is a window of tokens around the target. In our experiments we typically use a window of 7 on each side, where this includes **space tokens**, so that the effective window size in terms of non-space tokens is 3 on each side. The target word in our example is *beginning* and the phonemic subsequence corresponding to the target is highlighted in blue.

Returning to the concept of attention, applied to the problem of logography, one can consider that in order to predict the spelling of a given word, the model has to attend to various portions of the input. If the system is a simple phonographic system with essentially no logography, then the trained model ought to be able to find what it needs by attending just to the pronunciation of the target word. In a more logographic system, however, the model would need to look more broadly at the context because how one writes the word depends on what it means, not just on how it is pronounced. If one were to plot the attention matrix for a purely phonographic system, one would expect most or all of the attention activity to be confined to within the context of the target word; for a logographic system, one would expect the attention to be more spread out

9 http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

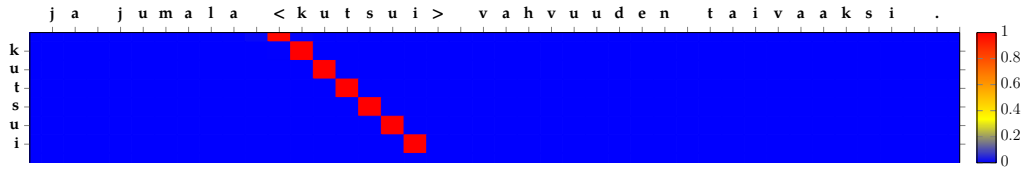


Figure 7
Attention matrix involved in spelling the Finnish word *kutsui* 'called'. The input (phonetic) sequence for the sentence is shown across the top of the plot, and the spelling of the target word is shown on the vertical axis. Note that in the plot itself the `<targ> ... </targ>` tags are reduced to just `<...>`. The active portion of the matrix—red—is almost entirely within the target word.

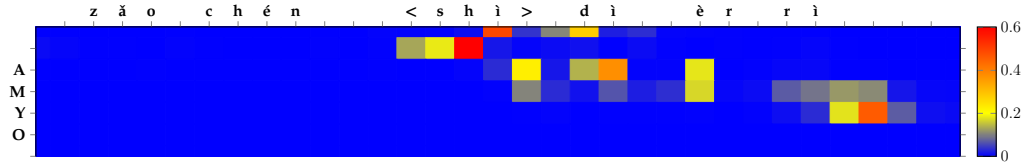


Figure 8
Attention matrix involved in spelling the Cangjie-encoded Chinese morpheme 是 (Cangjie AMYO) *shì* 'be'. (See Section 6 for details on encodings used for Chinese.) The input (phonetic) sequence for the sentence is shown across the top of the plot, and the spelling of the target word is shown on the vertical axis. The active portion of the matrix is spread out across much of the sentence.

across the context. This is illustrated in figures 7 and 8 for Finnish words versus Chinese morphemes (single characters).

5. Computational Measures of Logography

In this section we provide three classes of measures of the **distinct homophones** notion of logography: two measures based on attention; two simple lexical-based measures that can be viewed as the baseline measures that get at the same notion; and two measures based on n-gram entropy.

5.1 Attention-based Measures of Logography

To recap, in an ideal phonographic system, one would know how to spell a word based purely on its pronunciation. Conversely, in a highly logographic system, knowing how to spell a word would depend on the context. As we have seen in the previous section, in the former case, one would expect that an attention model would focus its attention mostly within the target word, since it should be able to determine purely on the basis of the phonemic representation of the word, what the spelling should be. Conversely, in the latter, logographic, case some portion of the attention would be spread out across the context containing the target word, since the system would need to look beyond the word to decide how it should be written.

One way to measure the *attention spread* would be to sum the activation over the attention matrix, then zero out the rectangle covering the target word's pronunciation and its spelling, by computing the Hadamard product (Horn and Johnson 2012) of the attention matrix with a mask matrix M whose entries i, j are 0 if $0 \leq i < k$, where k is the length of the target word's pronunciation, and $m \leq j \leq n$, where m is the left edge of

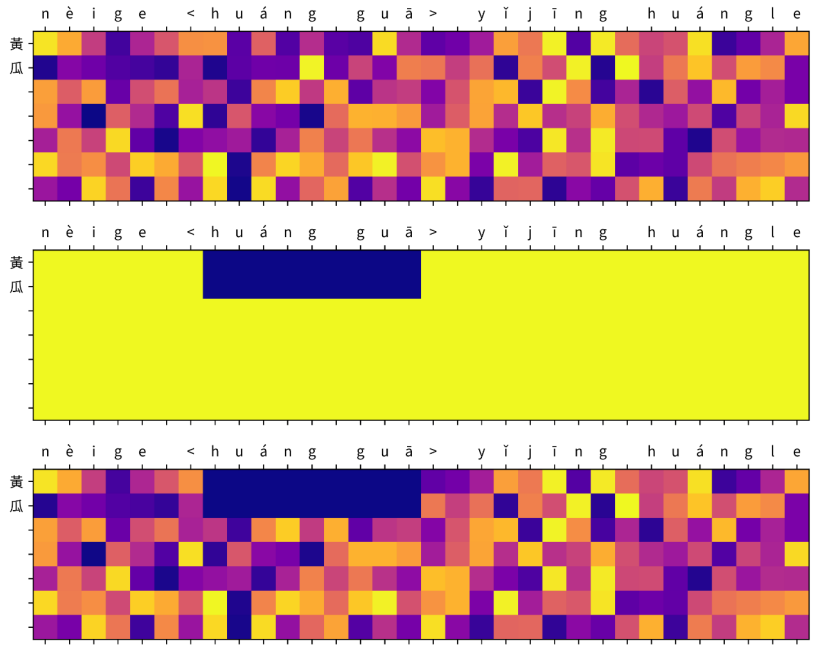


Figure 9
Illustration of the attention-based spread measure. Top: A random attention matrix. Middle: The zero mask for the target word. Bottom: The Hadamard product of the mask with the attention matrix.

the target word and n the right edge; and 1 otherwise. Then sum the resulting matrix, and divide it by the sum computed in the first step. Mathematically we define the spread for word w, S_w as

$$S_w = \frac{\sum_{i,j}(M \circ A)_{i,j}}{\sum_{i,j} A_{i,j}} \tag{7}$$

where A is the attention matrix and $M \circ A$ is the Hadamard product of the mask with the attention matrix. To illustrate, consider Figure 9 for a Chinese sentence *nèige <huáng guā> yǐjīng huángle*, 'that cucumber has already turned yellow', where the target word is *huángguā* 'cucumber', and the output spelling is 黄瓜. In this example for our attention matrix we compute a random matrix shown at the top. The zero mask is shown in the middle, and the result of applying that mask to the attention matrix is shown at the bottom. S is computed from the first and last matrices as defined above.

The measure of S for the system is then simply the mean value of S_w over the entire test set. This can be further broken down into S_{token}, which simply computes the mean over S_w for the corpus, versus S_{type}, which computes the mean within all instances of each w, and then computes the mean over these:

$$S_{\text{token}} = \frac{\sum_w S_w}{N} \quad \text{and} \quad S_{\text{type}} = \frac{\sum_v \frac{\sum_{w \in v} S_w}{|v|}}{V} \tag{8}$$

where N is the size of the corpus, V is the size of the vocabulary, w is a particular word instance, v is a particular word type, and $|v|$ is the number of instances of type v in the corpus. A purely phonographic system would have S values close to 0, whereas a highly logographic system should have S values close to 1. The intuition behind the type versus token-based measures is of course that the token-based measure will yield a value that is higher if it happens that the language has a few very frequent pronunciations that also happen to be very ambiguous as to what word they correspond to; whereas the type-based measure will be more balanced in that it will not overweight frequent terms. However, as we shall see below, S_{type} and S_{token} yield very similar rankings for our language samples, and so we will generally refer to these measures collectively as S where there is no need to distinguish the two measures.

Because the attention model obviously will make mistakes in prediction, in the experiments reported below we compute S over *only the correctly predicted forms*, as opposed to all predicted forms. This accords with intuition about how people process logographic forms: If a person who in principle knows the ways in which a spoken form could be written, nonetheless gets a particular instance wrong, it is likely because they are not gleaning the correct information from the context that would allow them to derive the correct spelling. If they spell the form correctly, then it is at least reasonable to assume that they may have made appropriate use of contextual information. Since the appropriate use of contextual information is key to the **distinct homophones** notion of logography, it seems therefore reasonable to concentrate on the cases where the system gets the spelling right.

Finally, before we consider other, non-neural measures of logography, we note that, as the reader is of course well aware, the attention mechanism has been the topic of intense scrutiny in recent work, with many researchers investigating what linguistic information is encoded in attention-based models and where and how it is encoded. This has been a particularly important area of research with deep attention models such as BERT (Voita et al. 2019; Clark et al. 2019; Rogers, Kovaleva, and Rumshisky 2021; Ravishankar et al. 2021), but there is also interest in exploring such questions for simpler RNNs—see, for instance, Silfverberg et al. (2021), who ask whether the encoder in an attentional bidirectional LSTM model encodes abstract phonological information. In many ways our use of attention here is a lot simpler: We are merely asking whether in a more logographic writing system, the system pays attention to material beyond the target word and to what degree. We are therefore using attention as a proxy for what we assume human spellers must consider when they decide how to spell a word in context.

5.2 Simple Lexical Measures of Logography

A much simpler measure of logography would just compute the mean of the number of spellings for a given pronunciation found in a dictionary, or corpus. This has both a *type* and *token* interpretation:

$$L_{\text{type}} = \frac{1}{|D|}\sum_{p\in D}|s(p)| \quad \text{and} \quad L_{\text{token}} = \frac{1}{|C|}\sum_{p\in C} c(p)|s(p)| \tag{9}$$

In the equation for L_{type}, D is a dictionary—which can be derived simply by compiling the set of pronunciations and their spellings from one's corpus, $s(p)$ is the set of spellings for pronunciation p for each p in D. For L_{token}, C is the corpus, $s(p)$ is again the set of spellings for each pronunciation p, and $c(p)$ is the total count of each p. In either case a

value close to 1 is an indication of a phonographic system, whereas a value significantly above 1 reflects a logographic system. While these definitions of L are intuitive, we will argue that the attention-based measure previously introduced results in a more satisfactory metric.

5.3 Entropic Measures

Another possible measure of logography is based on the information-theoretic concept of entropy introduced by Shannon (1948, 1951). In a writing system that is more logographic, one would expect that the amount of information carried by the written side would be higher than that carried by the spoken side simply because the written form of the language encodes information that is not present in the sequence of phonemes. Conversely, the entropy for the written form should be lower than the entropy for the spoken form. Thus, if one compared the entropy of a written token given the written context $H(w_i|w_1 \dots w_{i-1})$ with that of the same tokens in their phonological form $H(p_i|p_1 \dots p_{i-1})$, we would expect that the following relation for an entropic measure E_{token} would hold for a logographic writing system:

$$E_{\text{token}} = H(w_i|w_1 \dots w_{i-1}) - H(p_i|p_1 \dots p_{i-1}) < 0 \tag{10}$$

As we shall see below, this in fact seems to be true for Chinese and Japanese in the various encodings we considered, but seems not to be terribly useful for ranking other languages.[10]

In the experiments below we use the OpenGrm N-Gram toolkit (Roark et al. 2012) to build separate written and pronunciation-based bigram token models represented as weighted automata, denoted $\mathcal{W}$ and $\mathcal{P}$, on the training data.[11] Given the held out parallel test data consisting of N tokens for the written ($\mathcal{W}_C$) and pronunciation ($\mathcal{P}_C$) sides we can attempt to formalize the measure E_{token} from Equation (10) via the concept of corpus cross-entropy (Jurafsky and Martin 2009) defined between the models and the corresponding test data:

$$E_{\text{token}} = H(\mathcal{W}_C, \mathcal{W}) - H(\mathcal{P}_C, \mathcal{P}) = \frac{1}{N}\left(\sum_{p \in \mathcal{P}_C} \log P_{\mathcal{P}}(p) - \sum_{w \in \mathcal{W}_C} \log P_{\mathcal{W}}(w)\right) \tag{11}$$

where the test data is treated as the true distribution.[12]

Given the three models for the written ($\mathcal{W}$), pronunciation ($\mathcal{P}$), and joint written/pronunciation ($\mathcal{J}$) forms, it is also possible to define information-theoretic entropy-based measures on types, rather than tokens. Since for a particular model the states of the weighted automaton represent the probability distribution of word types conditioned on their respective histories (Roark, Allauzen, and Riley 2013), the type-specific

10 As one of our readers observes, this result might be taken as evidence that Chinese and Japanese writing is in fact in a different category from other languages rather than merely being at one end of a spectrum. But as we shall see below in any case (see Figure 10) it is not that the entropic measures suggest a categorical split, merely that only for Chinese and Japanese does E_{token} fall below zero.

11 These are mixture models interpolated with the unigrams using Witten and Bell (1991) smoothing.

12 It is worth noting that this definition corresponds to a particular case of token-based conditional entropy $H(\mathcal{W}|\mathcal{P})$ defined via joint entropy $H(\mathcal{P}, \mathcal{W})$ as $H(\mathcal{P}, \mathcal{W}) - H(\mathcal{P})$ (MacKay 2003, page 138), where the written form completely determines the pronunciation, in other words $H(\mathcal{P}, \mathcal{W}) = H(\mathcal{W})$. In reality, however, this assumption rarely holds exactly.

measure can make use of state probabilities. In particular, we investigate the measure defined via the concept of mutual information (MacKay 2003, page 139) between the written and pronunciation distributions

$$E_{\text{type}} = I(\mathcal{P}, \mathcal{W}) = H(\mathcal{W}) - H(\mathcal{W}|\mathcal{P}) = H(\mathcal{P}) + H(\mathcal{W}) - H(\mathcal{P}, \mathcal{W}) \tag{12}$$

The individual entropies are computed using marginal and joint state distributions for models $\mathcal{P}$, $\mathcal{W}$, and $\mathcal{J}$, where $H(\mathcal{P}, \mathcal{W}) = H(\mathcal{J})$. In general, given a model $\mathcal{M}$, its model state-based entropy is given by $H(\mathcal{M}) = -\sum_{s \in \mathcal{M}} P_{\mathcal{M}}(s)\log(P_{\mathcal{M}}(s))$, where individual states s are treated as discrete outcomes (MacKay 2003, page 32). The measure E_{type} can be interpreted as average reduction in uncertainty about the written form $\mathcal{W}$ which results from discovering the pronunciation $\mathcal{P}$. When comparing writing systems we expect the mutual information to decrease the more logographic the systems become.

In preparing the data for the above computations, sentences containing words with null pronunciations (see below in Section 6.1 on why this sometimes occurred) were removed. One issue to bear in mind with the entropic measure E_{token} is that because it compares the entropy of the written and spoken forms, the measure is sensitive to how accurately the two sides reflect the actual situation in the language. The written side can be taken as given (modulo the choice of tokenization), but the pronunciations are automatically generated and in particular, as noted above, in general we perform no homograph disambiguation. This will inevitably make the pronunciations of the input text less variable than would have been the case if the pronunciations were completely accurate. In what follows, we will collectively refer to the entropic measures from Equations (11) and (12) as E when there is no need to distinguish them.

6. Experiments

In this section we study the behavior of various logography measures introduced so far. The details of the Bible corpus and the data preparation methods used for the main experiments are described in Section 6.1. The details of the default Bahdanau neural attention architecture are provided in Section 6.2 and the results of the main body of experiments are presented in Section 6.3. Section 6.4 describes the performance of neural logography measures for selected languages trained on alternative data from Wikipedia. In Section 6.5 we study how using higher-quality data derived using state-of-the-art pronunciation extraction pipelines affects our results. Section 6.6 describes our investigation of low-resource scenarios for a bunch of languages, where the possibly low-quality data are derived using a rule-based grapheme-to-phoneme approach. Finally, in Section 6.7 we explore more sophisticated neural attention architectures (multihead self attention in transformer models and multistep attention in temporal convolution networks) using both Bible and Wikipedia data.

6.1 Data and Data Preparation

Data were collected from the Bible Corpus (Christodoulopoulos and Steedman 2015) for English, French, Russian, Swedish, Finnish, Korean, Chinese, and Japanese. Since the Hebrew bible in the Bible Corpus was undiacritized, whereas the original Hebrew (Old Testament) Bible is traditionally fully diacritized, for Hebrew we used instead the

Old Testament data from Mechon Mamre.[13] The languages chosen represent a sample of the spectrum from highly non-logographic systems, represented here by Finnish and Korean, to much more logographic systems like Chinese, Japanese, and English (per Sampson [1985] and also Sproat [2016]). Hebrew presents an interesting case because, as we describe below, we run the experiments with two versions, one with Biblical and the other with Modern pronunciation. Since Hebrew spelling is archaic, and Modern pronunciation coalesces many of the phonemes that were distinct in Biblical Hebrew (Berman 1997; Hornkohl 2019), we would expect Modern Hebrew to look more logographic than its ancient counterpart.

The set of verses was divided into training and testing by randomly choosing verses from both the Old and New Testaments, or in the case of Hebrew, the Old Testament only. Note that the train-test division was uniform across languages so that for example Genesis 1:1 was in the training set for all languages, and Genesis 1:8 was in the test set for all languages. See Table 4 for the corpus sizes. Each verse was tokenized to a list of tokens, each followed by their pronunciation. Thus, for example for Japanese, Genesis 1:3 appears as shown in Example 1, with each Japanese token followed by its pronunciation approximated by Romaji.

Example 1

神/kami は/wa 「/" 光/hikari あ/a れ/re 」/" と/to 言/i わ/wa れ/re
た/ta 。/. する/suru と/to 光/hikari が/ga あ/a っ/tsu た/ta 。/.

The details of processing differ for each language, as outlined directly below and as partially summarized in Table A.1.

English. Words were tokenized based on whitespace and punctuation, and rendered into ARPAbet pronunciations using the Pronouncing toolkit,[14] which simply looks up words in the CMU Pronouncing Dictionary.[15] This is obviously incomplete and any windowed region in which a word's pronunciation was not found was eliminated from the data. The pronunciation prediction also does no homograph disambiguation, so that the pronunciation is completely predictable by lexical lookup on the basis of the spelling for this set.

French. Words were tokenized based on whitespace and punctuation and pronounced using the lexicon developed by New et al. (2004).[16] As with English, windowed regions with unknown words were eliminated. Also, as with English, no homograph disambiguation was performed.

Russian. Words were tokenized based on whitespace and punctuation and pronounced using the WikiPron lexicon (Lee et al. 2020).[17] As with English, windowed regions with unknown words were eliminated. Also, as with English, no homograph disambiguation was performed.

13 http://www.mechon-mamre.org.
14 https://pypi.org/project/pronouncing.
15 The choice of the CMU Dictionary is motivated by the fact that while it is dated, and there are better options available, it has the advantage of being completely open source.
16 http://www.lexique.org/databases/Lexique383.
17 https://github.com/CUNY-CL/wikipron/blob/master/data/scrape/tsv/rus_cyrl_narrow.tsv.

Swedish. Words were tokenized as in English. The pronunciations were derived using the public domain Swedish lexicon hosted by the National Library of Norway (Nasjonalbiblioteket 2011).[18] This manually transcribed lexicon was originally developed by Nordisk Språkteknologi (NST), a Norwegian language technology company. The phonetic transcriptions are loosely based on the SAMPA standard (Wells 1997). Although Swedish is known to have a higher percentage of homographs than English (Hedlund, Pirkola, and Järvelin 2001), we perform no homograph disambiguation.

Finnish. Words were tokenized as in English. Because Finnish orthography is close to perfectly phonemic (Aro 2017), we simply used the identity mapping, with lower-casing of any capitalized letters, to derive the phonemic representation.

Hebrew. Words were tokenized as in English. The Hebrew Bible (Old Testament only) is fully diacritized, where the diacritics indicate vowels, quality of various consonants, and in some cases distinguish between completely separate phonemes (e.g., /s/ versus /ʃ/). The standard orthography, both Ancient and Modern, mostly eschews these diacritics, so we took the *written* form to be the completely *undiacritized* text. The fully diacritized text was then used to generate two forms of Hebrew pronunciation, namely, Biblical and Modern, using simple rules to produce IPA-like transcriptions:[19] Once the full diacritization of the word is known, the pronunciation of both Biblical and Modern Hebrew is mostly straightforward, the main complexity having to do with the treatment of schwa. The main difference between Biblical and Modern pronunciation resides in the collapse of various consonant distinctions. For example, in Modern Hebrew, both ⟨ת⟩and ⟨ט⟩ are pronounced /t/, whereas in Biblical Hebrew the former was /t/ or /θ/ depending on the context and the latter was an emphatic alveolar /tˤ/. Given that Hebrew spelling is conservative, the result is that Modern Hebrew is more logographic than Biblical Hebrew since in principle spelling distinctions are retained to distinguish words that are otherwise pronounced the same.

Korean. Words were tokenized as in English. Hangeul syllable code points were then converted to sequences of jamo (letters). Pronunciations were produced using the Ko-Pron project,[20] which produces transcriptions in Revised Romanization (Doll 2017). Thus the word 하나님 ('god') is rendered as ㅎ ㅏ ㄴ ㅏ ㄴ ㅣ ㅁ /hananim.

Chinese. Chinese data were produced in four forms, along two dimensions, the first dimension being character-based (i.e., morpheme-based) versus "word" based, and the second being whether to represent characters as simply their Unicode code points, or using their Cangjie encodings. Cangjie is a Chinese input system that is structural in that the coding loosely relates to the structure of the character in terms of traditional character components; it thus provides a decomposition of characters into a smaller set of somewhat sensible units. The Bible Corpus provides the Chinese Bible in two forms, unsegmented, and segmented into word-like units using a segmenter based on the Peking University segmentation standard (Yu 2002). Cangjie forms for both were produced using the Unicode Consortium's Unihan Dictionary[21] (Jenkins, Cook, and

18 http://www.nb.no/sbfil/leksikalske_databaser/leksikon/sv.leksikon.tar.gz.

19 https://github.com/google-research/google-research/tree/master/homophonous_logography/hebrew.

20 https://pypi.org/project/ko-pron.

21 https://unicode.org/Public/UNIDATA/Unihan.zip.

Table 4
Summary of the data sets for each of the languages/conditions. Note that Chinese tokenized input units are given as 'words' because in general the segmentation quality is very poor and therefore the units only loosely correspond to Chinese words. The Korean unit is listed as a phonological phrase, which includes words and additional particles, in Korean terminology referred to as *eojeol*.

Language	Tokens			Types	
	# Train	# Test	Type	# Train	# Test
English	713,721	176,259	word	7,863	5,232
French	749,359	185,389	word	16,571	9,648
Finnish	541,853	134,317	word	48,127	22,000
Russian	492,461	121,584	word	24,613	12,373
Swedish	515,230	128,035	word	15,156	8,875
Hebrew (Biblical)	277,657	86,014	word	38,225	17,040
Hebrew (Modern)	277,657	86,014	word	37,647	16,855
Korean (jamo)	378,565	94,136	phon. phrase	56,384	24,272
Chinese	822,317	204,558	morpheme	3,129	2,627
Chinese (Cangjie)	822,317	204,558	morpheme	3,127	2,626
Chinese (tokenized)	542,955	134,964	'word'	45,063	18,312
Chinese (tokenized, Cangjie)	542,955	134,964	'word'	45,060	18,312
Japanese	1,020,638	254,404	morpheme?	12,948	7,556
Japanese (Cangjie)	1,020,638	254,404	morpheme?	12,948	7,556

Lunde 2020). Finally, pinyin transcriptions were produced using the Pinyin project,[22] and these were used as pronunciations. The pinyin transcriptions include tone but, as with English, perform no homograph disambiguation. As examples of the four forms, in the basic configuration the expression for 'heaven and earth' would appear as "天/tiān 地/dì," and in the Cangjie encoding as "MK/tiān GPD/dì." In the tokenized version these two characters are tokenized together "天地/tiāndì," and in the Cangjie version appear as "MKGPD/tiāndì."

Japanese. Japanese was segmented using the KyTea project tools (Neubig and Mori 2010; Neubig, Nakata, and Mori 2011),[23] and converted to Romaji using the JPhones project.[24] The result of the segmentation is somewhat intermediate between doing no segmentation at all, and segmenting the text into linguistically sensible units such as *bunsetsu* (accentual phrases). That is, words seem to be segmented out as units, but particles are usually treated as a unit by themselves. This has the rather odd result that segmentations like あ/a っ/tsu た/ta ('it was') occur, as in the example from Genesis 1:3 above, whereas a more sensible segmentation would be あった/atta. In addition, a Cangjie version of the Japanese corpus was produced by replacing all kanji by their Cangjie code, where available: Cangjie is not used for inputting Japanese text, but it serves the same function as in our treatment of Chinese as providing a structurally motivated encoding of the Chinese characters. The above discussion is summarized in Table A.1 in the Appendix.

Corpus sizes for the various conditions are shown in Table 4. For each corpus division we list the number of written-spoken tokens, and the number of written-spoken (unique) types.

22 https://pypi.org/project/pinyin/.
23 http://www.phontron.com/kytea/.
24 https://github.com/JRMeyer/jphones.

The use of automatic tools to create pronunciations and (for some languages) tokenization for the corpora obviously has drawbacks; we discuss this issue further below, especially in Section 6.5.1, where we compare the results for Japanese with those from a higher quality Japanese tokenization and pronunciation system, and in Section 6.5.2 where we report a similar experiment for English. However we would like to clear up one possible misconception up front. One of the reviewers expressed concern about the methodology, asking how it would be possible to transcribe a text phonetically without already knowing how logographic a system is. But bear in mind that the measure of logography that we are investigating here is the phoneme-to-grapheme direction, whereas an automatic pronunciation system converts in the other direction, from graphemes to phonemes. It would at least be theoretically possible to have a system where it is almost always straightforward to phonetically transcribe a text automatically, because each written symbol has only one pronunciation; but where determining how to *spell* a given phoneme sequence requires one to consider the context. Chinese is, in fact, almost such a system, since most characters have only one pronunciation, or at least only one that is at all common; whereas determining the written form of a spoken word often requires broader context.

Also, it is necessarily the case that different languages have required different processing schemes: different tools and lexical resources with different quality; in some languages there is the necessity of doing word segmentation and indeed deciding what should count as a token. Could these choices affect the results? Certainly they can, and as we show below it makes quite a bit of difference, for example, in how we segment Chinese text: We argue in Section 6.3 that this difference actually makes sense if one considers what the **distinct homophones** notion of logography must mean operationally. On the other hand, our Modern and Biblical Hebrew processing was for all intents and purposes identical, yet yielded different results and, as we argue below, in the expected direction. On balance, while clearly one can expect that processing details will affect the results, we have no reason to believe this situation is any different for our work than it is for any research involving comparative multilingual NLP.

6.2 Default Neural Architecture Details

As discussed in Section 4, we utilize a neural sequence-to-sequence model (Sutskever, Vinyals, and Le 2014), where the input side corresponds to the phonemes in a discrete phoneme sequence and the output side represents the discrete orthographic symbol sequence. Recall from Section 4 that our input context is a window of 3 non-space tokens on each side of the target word. Our model is an instance of RNN encoder-decoder architecture with additive attention mechanism (Bahdanau, Cho, and Bengio 2015). The encoder embeds the inputs into a sequence of vectors in a continuous vector space, while the decoder component predicts the next symbol of the output, given the inputs and the previous outputs. Our model is implemented in TensorFlow (Abadi et al. 2016) with Keras abstractions (Géron 2019), and is derived in part from the TensorFlow Neural Machine Translation (NMT) tutorial.[25]

The encoder consists of an embedding layer (using a uniform initializer for the embedding matrix) that maps the inputs into continuous space with dimension $d = 256$, followed by a single recurrent layer of 256 gated recurrent units (GRUs) by Cho et al. (2014) with the following defaults: initialization by the Glorot and Bengio (2010),

25 https://www.tensorflow.org/tutorials/text/nmt_with_attention.

Table 5
Neural (S_{token}, S_{type}), lexical (L_{token}, L_{type}), and entropic (E_{token}, E_{type}) logography measures computed on the Bible corpora. Recall that lower values for S and L measures correspond to a lower degree of logography (inverse is true for E_{type}). The unexpectedly low values for some of the Chinese and Japanese encodings for L as opposed to the expected higher values for the same with the S measures suggests that on balance the neural attention-based metric is better at capturing the notion of logography. The fourth column gives the per-token spelling accuracy of the neural S model on the test data. For Russian and Swedish the † marker in the columns for the E measures indicates that the number is suspect because a very small number of training and testing verses (3,289 and 3,823, respectively) were left after removing verses that contained null pronunciations.

Language	Neural			Lexical		Entropic	
	S_{token}	S_{type}	Accuracy	L_{token}	L_{type}	E_{token}	E_{type}
Chinese	1.00	1.00	0.85	4.46	2.96	−0.12	7.86
Chinese (Cangjie)	0.74	0.71	0.87	4.45	2.96	−0.12	7.85
Chinese (tokenized)	0.55	0.37	0.89	2.10	1.05	−0.02	9.43
Chinese (tokenized, Cangjie)	0.51	0.32	0.78	2.10	1.05	−0.02	9.42
English	0.40	0.32	0.95	2.08	1.15	0.02	8.05
Finnish	0.19	0.12	0.96	1.43	1.05	0.02	10.10
French	0.57	0.36	0.89	3.10	1.68	0.14	8.24
Hebrew (Biblical)	0.65	0.50	0.94	1.06	1.04	0.06	9.18
Hebrew (Modern)	0.72	0.56	0.87	1.19	1.06	0.05	9.14
Japanese	0.97	0.88	0.94	7.19	1.25	−0.05	7.38
Japanese (Cangjie)	0.88	0.65	0.92	7.19	1.25	−0.06	7.38
Korean (jamo)	0.26	0.21	0.96	1.06	1.01	0.00	12.21
Russian	0.46	0.29	0.89	1.58	1.10	†0.12	†8.87
Swedish	0.35	0.20	0.90	1.13	1.01	†0.01	†8.95

hyperbolic tangent tanh activation function and a sigmoid for the recurrent activation function.[26] The decoder consists of a single GRU layer, identically configured to the one in the encoder, that takes its additional inputs as context vectors from the additive attention mechanism by Bahdanau, Cho, and Bengio (2015), followed by the usual feed-forward logits layer. We optimize the sparse categorical cross-entropy function using the Adam optimizer (Kingma and Ba 2014) with default hyper-parameters: initial learning rate $\alpha = 0.001$, exponential decay rates for the first and second moments $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. We use a batch size of 256.

6.3 Results on the Bible Corpus

The attention model was trained in the pronunciation-to-spelling direction for each of the languages/conditions until reasonable performance (low loss) was obtained on the training set. We found that 5 epochs was generally sufficient for the training to converge. The models were then evaluated on the held out test verses, omitting trivial predictions such as the "pronunciation" of punctuation symbols. Neural model accuracy for the different languages is shown in the fourth column of Table 5, where the language configurations are sorted in alphabetic order.

The neural logography measures S_{token} and S_{type} were computed as described above and the lexical measures L_{type} and L_{token} were computed by compiling a dictionary of words and their pronunciations from the entire corpus for each language/condition. The entropic measures E_{token} and E_{type} were computed as described in Section 5.3. The

26 We use no dropout (Srivastava et al. 2014) or regularization.

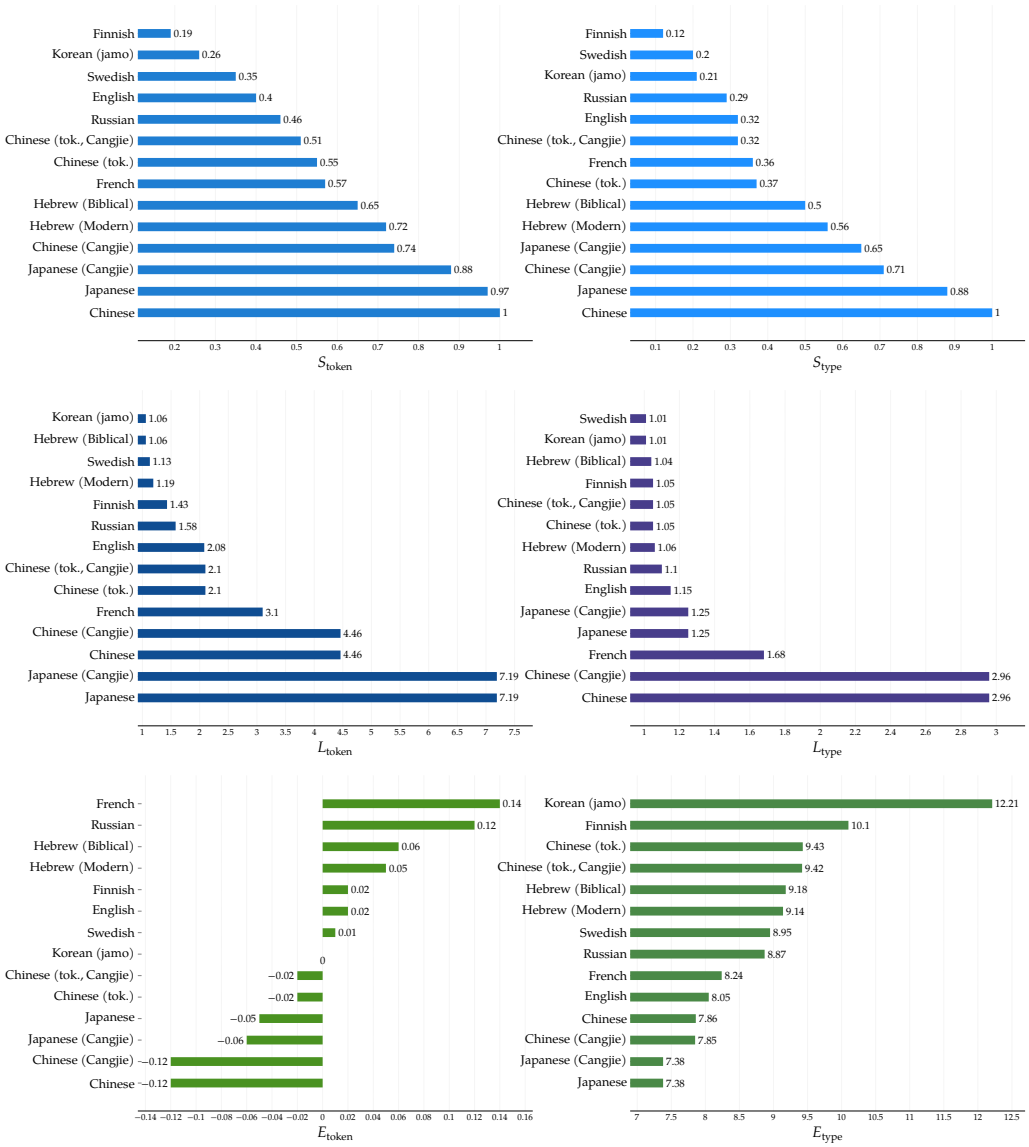


Figure 10
Language ranking (in increasing order of logography) according to S_{token} and S_{type} (top row), L_{token} and L_{type} (middle row), and E_{token} and E_{type} (bottom row).

resulting measures are given in columns 2–3, 5–6, and 7–8, respectively, of Table 5. To compare the six measures, we first sorted the S and L measures in the increasing order of logography (recall that lower values for these measures correspond to a lower degree of logography). Because E_{type} is defined via the concept of mutual information, which decreases with a higher degree of logography, we sort this measure in descending order. A similar sorting order is followed for E_{token}, which is based on the entropy difference. The resulting rankings are shown in Figure 10.

Note that there is not a great deal of difference between the token- versus type-based S measures. The ranking of Chinese (Cangjie) and Japanese (Cangjie) are reversed, and

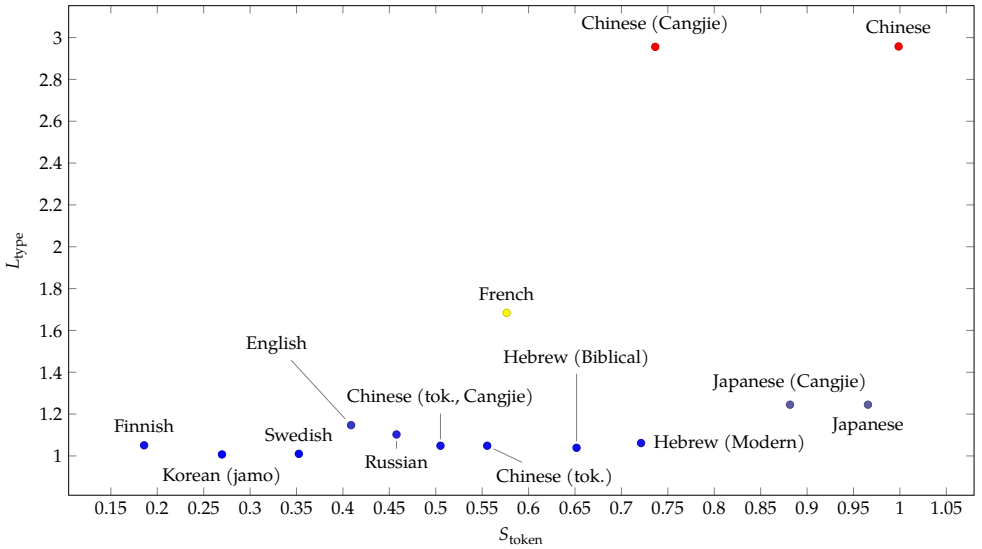


Figure 11
S_{token} from the attention model versus lexical ambiguity L_{type}.

similarly for French and Chinese (tokenized), and Russian and English, but these are all local perturbations. The E_{token} measure seems to work well for Chinese and Japanese in that it places them as solidly logographic, but for other writing systems the ranking does not seem to make a great deal of sense. The ranking for E_{type} is sensible in placing the non-tokenized Chinese and Japanese, as well as French and English, among the highly logographic systems, but misplaces tokenized versions of Chinese and assigns Swedish a more logographic rank than both versions of Hebrew, which in turn bizarrely ranks higher than tokenized Chinese.

Figure 11 plots S_{token} (horizontal axis) against L_{type} (vertical axis). While both L_{type} and S_{token} rank Chinese and Chinese (Cangjie) as highly logographic, L_{type} ranks French as being more logographic than Japanese. In contrast, the S_{token} measure ranks Chinese and Japanese as most logographic, with only tokenized Chinese being ranked lower—in this case lower than both varieties of Hebrew and French. Also, Finnish is the least logographic system according to S_{token} in contrast to L_{type}, which ranks it as more logographic than Hebrew, Korean, and Swedish. Interestingly, the estimates of both measures for Swedish support our hypothesis that logographically, unlike its Danish relative (Elbro 2006), Swedish is somewhat closer to Finnish than it is to English given its shallow and relatively uncomplicated orthography (van Daal and Wass 2017). On balance, S seems to accord more with intuition than L or E.

One thing that we note in the case of the Chinese single-character output (shown in the first row of Table 5), is that in this case the attention model seems to be waiting until it has read the whole input before it outputs its decision. This seems to be unique to this condition, and it obviously has the result that virtually all of the attention is outside the masked region. See Figure 12 for an example.[27]

27 This issue may also relate to the concerns expressed by Koehn and Knowles (2017), in particular in section 3.5, where they note problems with interpreting attention in terms of word alignment in neural MT systems. Generally though, how exactly attention corresponds to linguistic intuitions is less of a

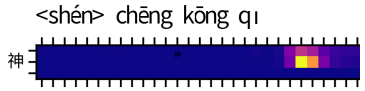


Figure 12
An example showing the attention for the output character 神 'God' in the Chinese single-character output condition. Note that the attention in this case is all past the end of the input.

For what it is worth, the S measures also seem to yield results that are more in accord with Rogers' intuition about the relative amount of logography (morphography) in various systems; see, again, Figure 2. Of the systems mentioned by Rogers and treated here, Rogers provides the following ranking from more to less logographic:

Japanese > Chinese > English > Russian > Korean > Finnish

This mostly accords with the ordering assigned by S_{token}, in that the majority of the encodings considered for Chinese are considered less logographic than Japanese, and the rest of the ordering is the same, with the exception of the placement of English relative to Russian:

Japanese > Chinese > Russian > English > Korean > Finnish

With S_{type}, the same ordering is obtained, except that now English and Russian are ranked according to Rogers' intuitions:

Japanese > Chinese > English > Russian > Korean > Finnish

In contrast, both L measures rank Korean as less logographic than Finnish, which seems counterintuitive. Returning to the S measures, the main exception in both cases is Hebrew, which Rogers would treat under West Semitic, and which turns out to measure as much more logographic than his original scheme implied.

The difference for all measures for the various Chinese conditions, in particular the tokenized and non-tokenized versions, underscores an important point: How logographic a system is depends on what one is trying to spell. Each syllable in Chinese may be rendered by a variety of different Chinese characters, depending on which morpheme is intended. So the system seems highly logographic from that point of view. If one now expands the scope to involve the spelling of words rather than single morphemes, the amount of ambiguity is reduced. It follows then that in order to resolve how to spell a given phonological word, one will often have to look outside that word far less than if one is trying to spell a given morpheme. Thus while the syllable *dì* may correspond to any of 6 different characters in the Bible corpus, the pair of syllables *tiāndì* has a unique spelling 天地 'heaven and earth'. So, in order to spell that syllable sequence, one does not have to look further in the context than the two syllables themselves.

concern for our work, because we are only concerned with whether the attention falls largely inside, or outside, a fixed rectangle in the attention space.

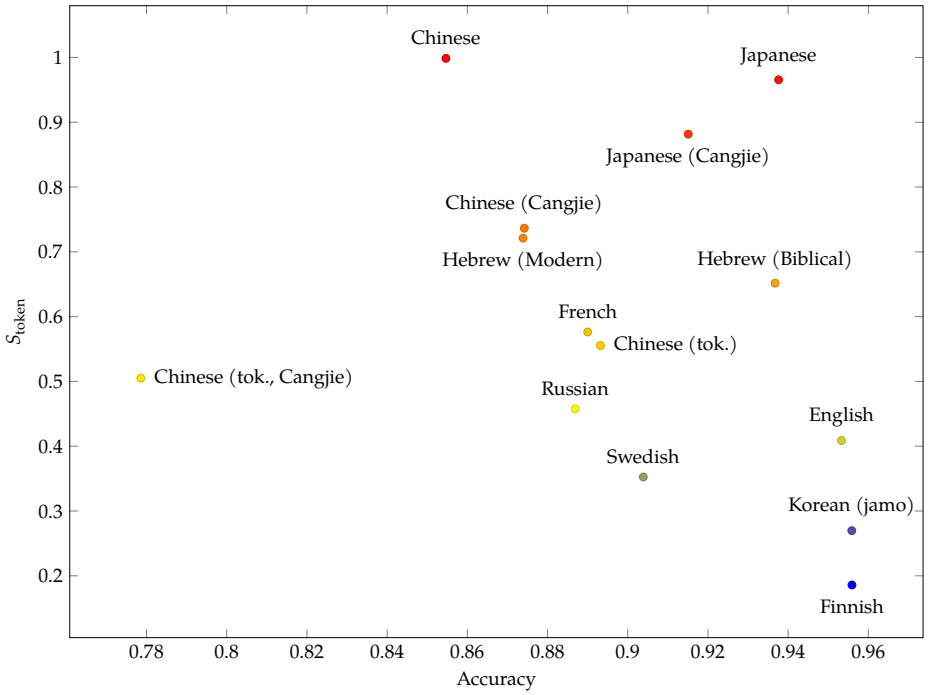


Figure 13
No correlation between accuracy and the spread S_{token} ($R^2 = 0.08$).

Finally, one concern is that there is some variation in the accuracy of the trained neural models. Could that be affecting our results? Figure 13 attempts to address that issue by showing that there is no correlation between accuracy and S_{token} ($R^2 = 0.08$), suggesting that the differing quality of the different trained models cannot explain the differences in the spread measure. Computing the same correlation for S_{type} and accuracy yields an even smaller $R^2(0.04)$.

6.4 Comparison for Selected Languages on Wikipedia

In order to compare our results across different text genres, we selected four languages that in our experiments on the Bible corpus were representative of minimal logography (Finnish), intermediate amounts of logography (Korean and English), and a high degree of logography (Japanese). We extracted sentences from Wikipedia for each of these languages, and filtered for clean sentences in that there were only language-appropriate graphemes, spaces, and punctuation (no numbers or other text normalization issues). In particular, for English and Finnish we only allowed letters, spaces, and punctuation; for Korean only *hangeul*, spaces, and punctuation; and for Japanese only *kanji*, *hiragana*, *katakana*, spaces, and punctuation. We also filtered to remove sentences more than 100 characters long. We made no attempt to select sentences that were parallel or even comparable.

We processed the Wikipedia data with the same tools as used for the Bible, and selected 1 million tokens of each language, selecting 90% for training and the remaining 10% for evaluation. We then trained our baseline neural model as before, and evaluated on the held-out data for S_{token} and S_{type}. Results are shown in Table 6, where the counts

Table 6
Neural (S_{token}, S_{type}) logography measures for selected languages on Wikipedia.

Language	Types		Neural		
	# Train	# Test	S_{token}	S_{type}	Accuracy
English	36,417	15,136	0.49	0.35	0.90
Finnish	145,146	36,409	0.18	0.12	0.90
Japanese	27,150	12,587	0.95	0.81	0.91
Korean (jamo)	150,145	42,524	0.28	0.23	0.90

of unique types in the training and evaluation sets are shown in the second and third columns, respectively. As can be seen, the results are not largely different from what we reported for the Bible in Table 5. Overall, accuracies are a bit lower, reflecting the wider range of vocabulary in Wikipedia. But the S values are only somewhat different, with the exception of Japanese, where S_{type} is 0.81, compared to 0.88 for the Bible corpus.

These results largely confirm our results from the Bible corpus for the rankings of these languages. Although a Bible translation is perhaps not optimal as a corpus for measuring the full range of logography in a language, it seems to be a good proxy if other data are not available.

6.5 Experiments with High-quality Data

6.5.1 Japanese. Although the experiments reported above are informative, as we have noted the results are limited by the quality of the phonetic transcriptions. In particular, most of the systems we report on do not include *homograph* disambiguation, which will inevitably bias to some extent our estimate of the level of ambiguity for some *homophones*. Also, in some languages, such as Japanese, the tokenization is rather poor.

We therefore ran an experiment on the Japanese Bible data, automatically transcribed with a high-quality text normalization system. Specifically, we used the Google RoadRuNNer neural text normalization system (Zhang et al. 2019), trained on about 700 million tokens of Japanese text, mostly from Wikipedia, but also a small amount of data from internal sources. The text itself had been previously verbalized using an internal Japanese text-to-speech language analysis component. The system was trained to provide pronunciations for all tokens. For Japanese, the RoadRuNNer system has a roughly 1.5% word error rate on held out hand-corrected data. Since for internal purposes pronunciations are rendered in *hiragana*, we used the Romkan package[28] to convert the pronunciations to romaji.

On the test data, the accuracy of the model was 0.91, and the S_{token} measure was 0.70, which is lower than what is found in Table 5. This can be attributed to the fact that for the original Japanese corpus, the mean written token length was 1.38 characters, whereas for the RoadRuNNer processed corpus, the mean word length is 1.89 characters. This, in turn, is due to RoadRuNNer yielding more sensible tokenization than the KyTea tools used in the corpus processing discussed previously. Once again, how logographic a system seems to be under the **distinct homophones** measure depends upon what one takes the tokens of interest to be.

In a similar vein, by way of example, the L_{type} score is 1.1 (compared to 1.25 for Japanese in Table 5) and the entropic E_{token} score is -0.02 (compared to -0.05 in Table 5). Both of these render Japanese with the longer tokens less logographic under the **distinct homophones** measure.

28 `https://pypi.org/project/romkan/`.

6.5.2 English. Similar to the experiment for Japanese reported in Section 6.5.1, we ran an experiment with English, using the Bible data, and passing it through the linguistic processing component of the Google US English text-to-speech system, part of the Google Assistant. This component includes the Kestrel text normalization system (Ebden and Sproat 2014), a very large carefully curated pronunciation lexicon, machine learning to derive pronunciations for words not found in the lexicon (Rao et al. 2015), and a homograph disambiguation component that handles hundreds of homographs (Gorman, Mazovetskiy, and Nikolaev 2018).[29]

The accuracy of the neural model was 0.96, similar to that reported in Table 5. S_{token} was 0.41 and S_{type} was 0.29, again similar to the previously reported results. This in turn suggests that our method will work as long as one can get reasonable overall pronunciations for a language, and as long as different tokenizations do not come into play. Note that, unlike the case of Japanese, the different processing of the English text did not result in a different tokenization, and this seems to be the main cause of the difference in Japanese reported in the previous section.

6.6 Data from Additional Languages using Epitran

In this experiment, we processed the Bibles from Christodoulopoulos and Steedman (2015) investigating the languages supported by Epitran grapheme-to-phoneme (G2P) conversion framework (Mortensen, Dalmia, and Littell 2018). Some of these languages, like Swedish, were evaluated above using hand-curated pronunciation dictionaries, while the others, like Shona (van de Velde et al. 2019) and Somali (Saeed 1999), are low-resource and lack reliable pronunciation resources. The motivation behind this is to investigate the performance of logography measures in low-resource language scenarios where pronunciations derived by automatic means, with Epitran or other approaches (Deri and Knight 2016; Novak, Minematsu, and Hirose 2016; Kirchhoff et al. 2018), are the only option available.[30,31]

The results are shown in Table 7 for the neural attention-based measures S and the lexical measures L. The same hyperparameters as in Section 6.2 were used for training the attention model with the exception of the number of epochs, which we set to 10: Epitran provides pronunciation predictions for all words, meaning that there are no lost examples unlike with some of the languages in our previous experiments with other open-source pronunciation tools; this results in larger amounts of training data for the attention model. The corresponding language rankings according to the attention-based measures S and lexical measures L are shown in Figure 14.

There are three languages (French, Russian, and Swedish) for which we can compare the S and L measures between the higher-quality pronunciations in Table 5 and the auto-generated ones. The results for both S measures accord with our intuition that the hand-curated lexicons naturally contain more pronunciation variation that on

29 See `https://github.com/google/WikipediaHomographData/blob/master/data/wordids.tsv` for a list of homographs covered.

30 Burmese and Thai are omitted from our set due to segmentation issues with the original data. Vietnamese is omitted because while tone is marked in standard Vietnamese orthography, tone is not supported by the Epitran pronunciation rules. Also among the omitted languages are Swahili, Ukrainian, and Zulu, for which only the New Testament portion of the Bible was available, making the amount of training data for the attention model significantly smaller than for other languages.

31 It is worth noting that in this study we use simple rule-based G2P methods, leaving more sophisticated state-of-the-art neural approaches to low-resource G2P using multilingual representations (Peters and Martins 2020; Zhao et al. 2020) for future work.

Table 7
Neural and lexical logography measures with pronunciations derived using Epitran. Neural model accuracy is reported per token. The prefix "S. A." in Spanish refers to pronunciation compromise between South American and Castilian dialects of Spanish adopted by the developers of Epitran.

Language	Training		Test (Neural)					Lexical	
	# Tokens	# Types	# Tokens	Accuracy	# Types	S_{token}	S_{type}	L_{token}	L_{type}
Amharic	332,197	76,512	159,092	0.973	27,851	0.2264	0.1820	1.0320	1.0028
Cebuano	693,146	25,847	335,985	0.963	11,058	0.3207	0.1771	1.7781	1.0761
Dutch	580,215	21,404	282,528	0.919	9,210	0.3542	0.1982	1.8632	1.0787
Farsi	537,752	38,037	261,287	0.963	14,451	0.3392	0.2172	1.0267	1.0050
French	580,403	24,234	280,557	0.870	7,714	0.5327	0.3435	3.7134	1.3596
German	558,747	20,626	270,213	0.925	7,944	0.3620	0.2179	1.9737	1.1289
Hindi	633,357	19,773	308,223	0.985	9,506	0.2550	0.1743	1.0543	1.0146
Hungarian	480,274	61,494	229,326	0.949	21,819	0.2585	0.1237	1.6233	1.0564
Indonesian	505,765	20,400	243,553	0.908	8,357	0.2513	0.1739	1.7966	1.1770
Marathi	510,964	48,096	245,707	0.983	19,613	0.1984	0.1411	1.0079	1.0018
Polish	488,441	40,902	226,455	0.959	15,991	0.3098	0.1774	1.7010	1.0932
Portuguese	560,116	28,677	270,904	0.947	11,284	0.3430	0.1845	1.9674	1.0919
Romanian	564,080	24,114	273,367	0.943	11,088	0.2888	0.1678	1.7241	1.1065
Russian	450,150	45,387	215,665	0.897	16,009	0.3217	0.1595	1.6182	1.0540
Shona	364,983	59,833	174,503	0.903	19,738	0.1616	0.1148	1.5545	1.0822
Somali	582,823	37,630	283,459	0.952	15,665	0.3049	0.1986	1.6129	1.0794
S. A. Spanish	573,958	27,105	277,761	0.947	11,016	0.6452	0.5396	2.3041	1.1147
Swedish	587,907	25,685	283,275	0.944	10,895	0.3575	0.1955	1.7326	1.0612
Tagalog	661,561	23,584	321,259	0.978	10,460	0.2982	0.1392	2.1257	1.0803
Telugu	354,095	82,033	170,049	0.990	30,213	0.1655	0.1126	1.0002	1.0000
Turkish	360,427	57,819	172,917	0.908	21,638	0.1778	0.1239	1.5369	1.1052
Xhosa	355,175	77,149	169,499	0.910	24,816	0.1904	0.1406	1.5523	1.0830

average results in a *higher degree of logography* than the corresponding configurations in Table 7. This is especially true for French and Russian, where the differences between S measures computed on hand-curated and auto-generated pronunciations, denoted as $(\Delta S_{\text{token}}, \Delta S_{\text{type}})$, are $(0.04, 0.02)$ for French and $(0.14, 0.13)$ for Russian, respectively. This difference is negligible for the S measures on Swedish, which indicates both the relatively high quality of Swedish pronunciation rules in Epitran compared with both French and Russian, and the relatively low degree of Swedish logography that allows more context-independent pronunciation rules to correctly determine the orthography. For the L measures things are a bit more mixed: For L_{token}, the Epitran pronunciations yield higher logography values for all three languages. For L_{type} Swedish counts as slightly more logographic according to the Epitran pronunciations, but Russian and French as less logographic.

The results for some of the languages in the ranking shown in Figure 14 may seem surprising. For example, Spanish is typically held up as a language that has a highly regular spelling system, so it seems odd that it ranks as highly logographic under both the S and L measures. But a moment's reflection will reveal that when people note this property of Spanish, what they are referring to is the fact that it is easy to know how to *pronounce* a word in Spanish, given the spelling. Indeed, Spanish orthography even goes so far as to indicate lexical stress, when that cannot be induced from the regular letter-to-sound correspondences of the system. But in order to know how to *spell* a particular word, one often has to know which word from a set of homophonous words are intended. Thus in our corpus, /sera/ can be spelled ⟨sera⟩ 'evening', ⟨será⟩ 'will be', or ⟨cera⟩ 'wax'. The first two arise since the Epitran pronunciations do not indicate stress, so one might discount these, but the first and third are valid homophones for

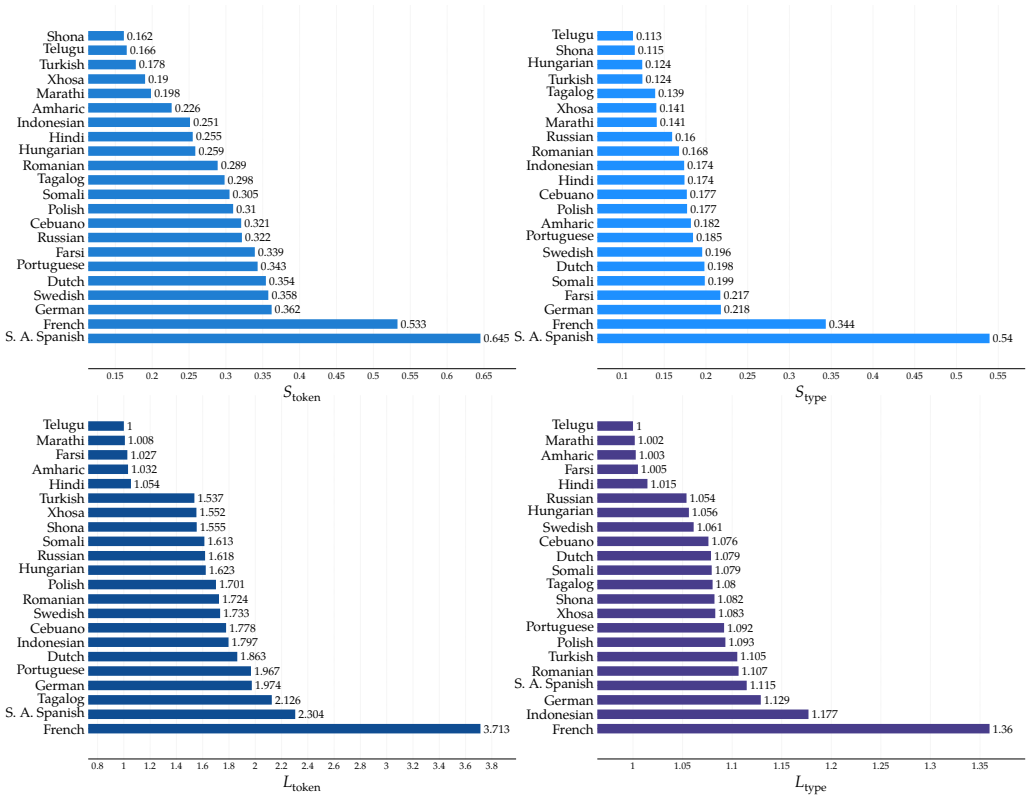


Figure 14
Language ranking with Epitran pronunciations (in increasing order of logography) according to neural (S_{token} and S_{type}, top) and lexical (L_{token} and L_{type}, bottom) measures. In comparing this plot with that in Figure 10, please note that the scales are different.

most dialects of Spanish. As another example consider /kaso/, which occurs as ⟨caso⟩ 'case' ⟨casó⟩ 'married' or ⟨cazó⟩ 'hunted'. There are in addition to these legitimate cases instances where what is apparently the same name is spelled differently in different portions of the corpus—for example, ⟨Jahaziel⟩, *Jaasiel* and *Jaaziel*, but one finds similar variation in, for example, the English and Finnish corpora. So, in summary, at least some of the choices the system is forced to make are legitimate, and lead to the conclusion that from the point of view of the **distinct homophones** notion of logography, Spanish is quite logographic.

For Indonesian, the L scores are relatively high, but this seems to be due to variation in *within word* capitalization. Thus for /tindakanmu/ we find ⟨TindakanMu⟩, ⟨Tindakanmu⟩, ⟨tindakanmu⟩, and ⟨tindakanMu⟩. The S measures seem to be less sensitive to this variation. We encounter a similar issue with other languages, such as Shona. Modern Shona has uncomplicated and rather rigid orthography (Magwa 2002) and the S_{token} measure firmly ranks it as the least logographic of all languages in the list. Similar to Indonesian, however, the L_{type} measure for Shona cannot distinguish between capitalization variants, where the spelling variants for /namwari/ ('by god') include ⟨naMwari⟩, ⟨Namwari⟩, and ⟨NaMwari⟩. Some other higher logographic L_{type} scores for other languages, such as Turkish, can be explained similarly.

6.7 Investigation of Alternative Neural Attention Architectures

So far our experiments have focused on one of the simplest neural sequence-to-sequence architectures, the encoder-decoder RNN network with attention described by Bahdanau, Cho, and Bengio (2015). In this section we investigate the performance of neural logography measures using some alternative neural models of attention that have been proposed since. Although the models under investigation are significantly different structurally, it is interesting to note that fundamentally they still form a bipartite architecture, going back to the work of Cho et al. (2014) and Sutskever, Vinyals, and Le (2014), equipped with some form of attention mechanism of varying complexity. We hypothesize that this organizing principle should reinforce our tentative neural view of logography (based on the S measures computed over the alignments between pronunciations and corresponding spellings) for these types of more sophisticated models as well.

6.7.1 Transformer Architecture. This type of model was introduced by Vaswani et al. (2017). Unlike other bipartite attention architectures for sequence-to-sequence modeling that utilize recurrent or convolution layers in their encoder and decoder components, the transformer model replaces these in both the encoder and the decoder with N stacked layers each consisting of a *multihead self attention mechanism* and feed-forward network. An additional multihead self attention layer in the decoder serves as an interface to the decoder receiving its input from the encoder. The multihead self attention is a key component of the transformer architecture. Informally, within each transformer layer instead of computing attention once, the multihead mechanism splits the inputs into h smaller portions and then computes the scaled dot-product attention over each subspace independently. The parallel attention outputs are then concatenated and linearly transformed using a feed-forward network. Since its inception this architecture has become a de facto standard modeling paradigm, its numerous flavors and generalizations achieving state-of-the-art results in many NLP tasks (Devlin et al. 2019; Dehghani et al. 2019; Dai et al. 2019; Kitaev, Kaiser, and Levskaya 2020).

Network Details. Our transformer model is based on one of the public-domain TensorFlow implementations of the original model by Vaswani et al. (2017).[32] We use $N = 4$ layers in both the encoder and the decoder, $h = 8$ parallel heads in the attention mechanism, set the size of the feed-forward layer to $d_{\text{ff}} = 512$, the dimension for model inputs and outputs, and the respective embeddings, $d_{\text{model}} = 128$. The rest of the model parameter details, such as normalization and optimization strategy, are as per Vaswani et al. (2017). We set the batch size to 256 and train the model for 15 epochs. We denote the thus constructed transformer configuration V.

Multihead Self Attention and Logography. The canonical transformer model by Vaswani et al. (2017) has three attention mechanisms: the self attentions in the encoder and decoder, and the decoder-encoder attention in the decoder that performs attention over the encoder stack outputs. It is the latter mechanism that is of primary interest to us for the purpose of computing the neural logography measures S over a matrix of attention weights between pronunciations and corresponding spellings. One difficulty that naturally arises in the transformer setting is how to select the appropriate representation

32 https://blog.tensorflow.org/2019/05/transformer-chatbot-tutorial-with-tensorflow-2.html.

of attention weights given multiple self attention heads. There has been an increased research focus on analyzing the behavior of attention mechanisms in various flavors of transformer models in order to understand the linguistic function of the attention and also improve model compression schemes (Clark et al. 2019; Michel, Levy, and Neubig 2019; Vig and Belinkov 2019; Voita et al. 2019; Behnke and Heafield 2020; Wang et al. 2020; Rogers, Kovaleva, and Rumshisky 2021). While in-depth investigation into the precise role the multiple attention heads play for logography is outside the scope of this work, we opt for a simple strategy whereby we inspect multiple attention heads in the top layer of the decoder-encoder attention block. To compute the attention weight matrix over the multiple heads we utilize two basic approaches: averaging all the heads (Voita et al. 2018) (denoted V_A) and choosing the maximum element from any of the heads (Tang, Sennrich, and Nivre 2018) (denoted V_M).

Results. The evaluation of the transformer model on the Bible corpus is shown in Table 8, where the results for transformer configurations V_A and V_M (computed using the same transformer model V, hence we show a single accuracy column) are shown alongside our core RNN model (denoted B) from the previous sections (the evaluation data are described in Table 4). As can be seen from the table, transformers are significantly more accurate on our task, outperforming the Bahdanau RNN configuration by 4.6% on average over the given 14 writing system configurations.

Examining the languages from the original Rogers' logography ranking from Section 6.3, the first three languages with the most degree of logography (Japanese > Chinese > English) that obey Rogers' intuition are consistently ranked by all four neural transformer measures. The language rankings for the remaining least logographic languages vary. The "averaging" token-based measure $S_{\text{tok}}^{V_A}$ swaps Finnish and Korean around, while its type-based counterpart $S_{\text{typ}}^{V_A}$ swaps around Korean and Russian. The ranking produced by measure $S_{\text{tok}}^{V_M}$ for languages in Rogers' list is identical to the ranking

Table 8
Neural (S_{token} and S_{type}) logography measures computed on the Bible corpora. The first measure (denoted B) corresponds to the RNN configuration from Table 5 and is copied here to ease the comparison. The rest of the measures were computed for a single transformer model V, where V_A and V_M measures were computed by combining over multiple heads in the last decoder block by averaging and choosing the maximum element, respectively.

Language	B			V_A			V_M	
	S_{tok}^{B}	S_{typ}^{B}	Acc.	$S_{\text{tok}}^{V_A}$	$S_{\text{typ}}^{V_A}$	Acc.	$S_{\text{tok}}^{V_M}$	$S_{\text{typ}}^{V_M}$
Chinese	1.00	1.00	0.85	0.87	0.84	0.95	0.80	0.74
Chinese (Cangjie)	0.74	0.71	0.87	0.79	0.76	0.94	0.65	0.61
Chinese (tok.)	0.55	0.37	0.89	0.90	0.81	0.93	0.91	0.82
Chinese (tok., Cangjie)	0.51	0.32	0.78	0.84	0.75	0.90	0.84	0.73
English	0.40	0.32	0.95	0.81	0.73	0.98	0.75	0.68
Finnish	0.19	0.12	0.96	0.34	0.24	0.98	0.37	0.26
French	0.57	0.36	0.89	0.52	0.35	0.94	0.51	0.33
Hebrew (Biblical)	0.65	0.50	0.94	0.44	0.30	0.96	0.39	0.28
Hebrew (Modern)	0.72	0.56	0.87	0.53	0.35	0.92	0.49	0.31
Japanese	0.97	0.88	0.94	0.94	0.85	0.96	0.92	0.83
Japanese (Cangjie)	0.88	0.65	0.92	0.90	0.83	0.96	0.90	0.84
Korean (jamo)	0.26	0.21	0.96	0.33	0.26	0.99	0.33	0.27
Russian	0.46	0.29	0.89	0.38	0.25	0.93	0.40	0.27
Swedish	0.35	0.20	0.90	0.43	0.27	0.91	0.47	0.33

Table 9
Neural (S_{token} and S_{type}) logography measures computed on the subsets of Wikipedia using transformer architecture V. Multihead attention strategies similar to Table 8.

Language	V_A			V_M	
	$S_{\text{tok}}^{V_A}$	$S_{\text{typ}}^{V_A}$	Acc.	$S_{\text{tok}}^{V_M}$	$S_{\text{typ}}^{V_M}$
English	0.53	0.42	0.94	0.53	0.42
Finnish	0.46	0.40	0.96	0.54	0.50
Japanese	0.86	0.75	0.92	0.84	0.73
Korean (jamo)	0.36	0.31	0.97	0.34	0.30

from $S_{\text{tok}}^{V_A}$. The measure $S_{\text{typ}}^{V_M}$ is the only transformer-based measure out of the four that nearly ranks the languages according to Rogers' intuition, however it cannot distinguish between Korean and Russian, which both receive an identical score of 0.27. In addition, all the four neural measures correctly identify modern Hebrew as more logographic than Biblical Hebrew, but place English as significantly more logographic than French, as opposed to the inverse ranking obtained with our core RNN-based neural measures S.

It is worth noting that the scores in each of the four transformer-based logography rankings appear to be slightly more "clumped," offering smaller resolution compared to the corresponding RNN-based scores. For example, Finnish is not as "purely" phonographic and Chinese not as "purely" logographic according to transformer measures (V) if we compare them to the RNN ones (B). This is most likely due to the difficulties in interpreting multiple self-attention transformer heads that require us to combine them in ways that introduce unnecessary noise.

Finally, we computed the neural logography measures using the transformer architecture trained on the Wikipedia subsets of the four languages. The corresponding results obtained with the RNNs are shown in Table 6 of Section 6.4. The neural S measures computed from attention averaged across attention heads (V_A) and element-wise maximum (V_M) are shown in Table 9. Similar to our experiments on the Bible corpus, the transformer architecture outperforms its RNN counterpart in terms of pronunciation-to-spelling accuracy by about 4.5% on average. In terms of logography rankings, the best configuration again corresponds to the rankings V_A computed from attention averaged across multiple self attention heads. We note that the V_M configuration is clearly worse, ranking Finnish more logographic than English. In contrast, both the token- and type-based measures $S_{\text{tok}}^{V_A}$ and $S_{\text{typ}}^{V_A}$ produce a ranking of the form Japanese $>$ English $>$ Finnish $>$ Korean, which is better than V_M and more acceptable, but is still different from the more intuitive corresponding RNN ranking from Table 6 that places Korean above Finnish.

6.7.2 Convolutional Architecture. In this type of model the recurrent units of the RNNs are replaced with temporal convolutions while retaining the original encoder-decoder structure. Prior to the emergence of now dominant transformer-based architectures, temporal convolution networks (TCNs) were shown to be competitive with RNNs in certain scenarios on several NLP tasks, such as language modeling (Bai, Kolter, and Koltun 2018) and neural machine translation (Kalchbrenner et al. 2016; Kaiser, Gomez, and Chollet 2018). Evaluating this type of model equipped with an attention mechanism on the logography task is interesting because this architecture is sufficiently different from both the RNNs and transformers.

The TCN architecture in this study (which we denote G) mostly follows the fully convolutional sequence-to-sequence attention-based architecture proposed by Gehring et al. (2017), with some minor modifications. Both the encoder and decoder consist of stacked blocks with each block consisting of a temporal convolution layer with residual connection from the input of each convolution to the block's output, followed by non-linearity implemented with gated linear units proposed by Dauphin et al. (2017). Our modifications to the original architecture include dilated convolutions (Yu and Koltun 2016; van den Oord et al. 2016), which are causal in the decoder, inspired by the ByteNet architecture by Kalchbrenner et al. (2016), and the positional embeddings in the encoder described by Devlin et al. (2019).

Multistep Attention. Each decoder layer in this architecture is equipped with the attention mechanism that is mostly similar to the RNN attention by Luong, Pham, and Manning (2015) and is computed using the current decoder state, the embedding of the target element, outputs from the last encoder block, and, unlike the traditional RNN attention, the encoder input embeddings (Gehring et al. 2017). The context vector computed using the attention mechanism and the corresponding residuals is used as an input for the next decoder block. Unlike the transformers considered earlier, the attention in each decoder layer has a single head, but one still has multiple potential alignments to consider when computing neural logography measures S. For the TCN architecture we investigate five ways of inspecting or combining the alignments given multiple layers of attention: averaging over all the layers (A), selecting the maximum element from the available attention scores (M), multiplying the attention scores (J), and simply selecting the attention in the decoder's bottom (B) or top (T) layers.

Network Details. Our TensorFlow implementation is based on the original implementation of Gehring et al. (2017) in the FAIRSEQ toolkit (Ott et al. 2019),[33] but our parameters differ. Both the encoder and the encoder components consist of two temporal convolution blocks with 256 hidden units. Because the gated linear units require the dimension of the inputs to be twice the size of the outputs, this amounts to 512 hidden units for the actual temporal convolution layers. We set the convolution kernel width $k = 3$ and exponential dilation rates to 1 (for the bottom layer) and 2 (for the top layer). Causal padding is used in the decoder convolution layers. The dimensions of input and target embeddings is 64. During training we apply dropout (Srivastava et al. 2014) with a rate of 0.1 to all the embedding and convolution layers. Our optimization strategy follows the work of Vaswani et al. (2017), while the rest of the parameters including the residual scaling factors are used as described by Gehring et al. (2017). As with transformers, we employ a batch size of 256 and train the models for 15 epochs.

Results. We trained and evaluated the TCN architecture G on the Bible corpus. The results are shown in Table 10 where, in addition to the usual accuracy metric, we evaluate the logography measures S using five different ways of inspecting the TCN multistep attention mechanism within the same model: Averaging over all the layers (A), choosing the maximum element from the available layers (M), choosing the top (T) or bottom (B) layer, and element-wise multiplication (J). We start by observing that in terms of pronunciation-to-spelling prediction accuracy, our TCN architecture performs nearly as well as the transformer architecture (see Table 8); the average performance

33 https://github.com/pytorch/fairseq/.

Table 10
Neural (S_{token} and S_{type}) logography measures computed on the Bible corpora using fully convolutional architecture (denoted G). Attention combination strategies: Averaging (A), choosing maximum (M), top layer (T), bottom layer (B), and multiplying (J).

Language	G^A		G^M		G^T		G^B		G^J		Acc.
	$S^{G^A}_{\text{tok}}$	$S^{G^A}_{\text{typ}}$	$S^{G^M}_{\text{tok}}$	$S^{G^M}_{\text{typ}}$	$S^{G^T}_{\text{tok}}$	$S^{G^T}_{\text{typ}}$	$S^{G^B}_{\text{tok}}$	$S^{G^B}_{\text{typ}}$	$S^{G^J}_{\text{tok}}$	$S^{G^J}_{\text{typ}}$	
Chinese	0.63	0.63	0.63	0.63	0.59	0.60	0.67	0.67	0.73	0.73	0.89
Chinese (Cangjie)	0.42	0.41	0.42	0.41	0.47	0.50	0.37	0.31	0.49	0.47	0.93
Chinese (tok.)	0.54	0.36	0.54	0.36	0.39	0.26	0.69	0.46	0.99	0.98	0.91
Chinese (tok., Cangjie)	0.39	0.31	0.40	0.31	0.47	0.35	0.31	0.26	0.29	0.20	0.92
English	0.30	0.21	0.30	0.22	0.38	0.27	0.21	0.16	0.32	0.23	0.99
Finnish	0.29	0.20	0.30	0.21	0.28	0.18	0.31	0.22	0.27	0.18	0.99
French	0.49	0.29	0.49	0.30	0.45	0.27	0.52	0.31	0.38	0.21	0.94
Hebrew (Biblical)	0.44	0.35	0.45	0.36	0.44	0.34	0.45	0.36	0.43	0.33	0.96
Hebrew (Modern)	0.41	0.31	0.41	0.32	0.43	0.33	0.38	0.30	0.41	0.30	0.93
Japanese	0.75	0.56	0.75	0.56	0.91	0.67	0.59	0.44	0.97	0.87	0.94
Japanese (Cangjie)	0.55	0.37	0.55	0.37	0.75	0.43	0.35	0.30	0.57	0.35	0.96
Korean (jamo)	0.24	0.19	0.25	0.20	0.28	0.22	0.19	0.16	0.20	0.16	0.99
Russian	0.34	0.23	0.35	0.24	0.39	0.25	0.29	0.22	0.38	0.24	0.93
Swedish	0.41	0.31	0.41	0.32	0.39	0.22	0.43	0.40	0.34	0.18	0.92

degradation being around 0.4% computed over 14 writing system configurations. As can be seen from Table 10, in terms of logography measures, all the configurations rank Japanese and Chinese as the most logographic, although the S measures computed using the bottom layer (G^B) and using the element-wise product of attentions (G^J) both rank Chinese solidly above Japanese. Furthermore, the measures computed from layer-wise average (G^A) and element-wise maximum (G^M) rank Russian as more logographic than English and place Finnish above Korean. Perhaps unsurprisingly, we find that the most satisfactory configuration is G^T, which corresponds to the S measures computed using Equation (8) (Section 5.1) from the attention in the top layer of the decoder. Out of the two measures, the type-based $S^{G^T}_{\text{typ}}$ matches Rogers' ranking and *mostly* corresponds to our intuition. The resulting ranking is not without oddities—similar to all other rankings in Table 10, the best configuration places Biblical Hebrew as more logographic than Modern Hebrew and places English and French on equal footing, something at which its token-based counterpart $S^{G^T}_{\text{tok}}$ does a better job.

We also evaluated the TCN architecture on Wikipedia subsets of the four languages, similar to the previously described experiments with the RNN (Table 6) and transformer architectures (Table 9). Evaluation results for token- and type-based logography measures S computed for five TCN configurations are shown in Table 11. With the exception of Japanese, the TCNs for other languages outperform the corresponding RNN configurations by at least 3% in terms of accuracy. Inspecting the logography measures, both token- and type-based S measures corresponding to layerwise averaging (G_A), layerwise maximum (G_M), and bottom decoder layer (G_B), as well as the type-based product (G_J) measure, are problematic because they misplace Finnish above English. Similar to our findings on the Bible corpus, the configuration G_T corresponding to the neural measures computed over the attention in the top decoder block produces ranking that makes most sense: Japanese $>$ English $>$ Finnish $>$ Korean. This mirrors the ranking produced by the best transformer configuration (shown in Table 9) and has a similar problem of placing Finnish as more logographic than Korean.

Table 11
Neural (S_{token} and S_{type}) logography measures computed on the Wikipedia subsets using fully convolutional architecture G. Attention strategies identical to Table 10.

Language	G^A		G^M		G^T		G^B		G^J		Acc.
	$S^{G^A}_{\text{tok}}$	$S^{G^A}_{\text{typ}}$	$S^{G^M}_{\text{tok}}$	$S^{G^M}_{\text{typ}}$	$S^{G^T}_{\text{tok}}$	$S^{G^T}_{\text{typ}}$	$S^{G^B}_{\text{tok}}$	$S^{G^B}_{\text{typ}}$	$S^{G^J}_{\text{tok}}$	$S^{G^J}_{\text{typ}}$	
English	0.35	0.25	0.35	0.26	0.38	0.26	0.32	0.25	0.29	0.18	0.93
Finnish	0.42	0.37	0.43	0.38	0.27	0.21	0.57	0.54	0.27	0.20	0.96
Japanese	0.46	0.29	0.46	0.29	0.51	0.29	0.41	0.28	0.56	0.44	0.87
Korean (jamo)	0.28	0.25	0.29	0.26	0.23	0.19	0.33	0.30	0.20	0.17	0.98

6.7.3 Summary of Additional Neural Experiments. In this section we have examined two further neural architectures, namely, transformers and temporal convolutional networks, both of which are more "state-of-the-art" than the simple attentional neural model introduced in Section 4 and used in the main experiments above. While the more sophisticated models perform better on the task of converting from spoken to written form, the logography measures that we derive from them are not clearly better than those derived from the simpler model. Part of the problem is that because these models have many more moving parts—the transformer in particular has multiple layers and heads of attention mechanism that one might consider—it is harder to know what representation one should be using in order to derive the measure we seek. For both transformers and the convolutional models we tried several different possible ways of extracting the information, but it is still possible that we missed a way that would have produced a better alignment to our previous results.

Ultimately, the advantage of our initial model for this task is its great simplicity: Because there is but one attention matrix, there is no question about what we should be looking at, and the computation is therefore straightforward. This observation is in line with the findings of the line of research that deals with NLP model interpretability and faithfulness which prefers simpler neural architectures (Wiegreffe and Pinter 2019; Moradi, Kambhatla, and Sarkar 2021).

7. Critique and Limitations

In Section 2 we presented two distinct ways in which a writing system could be considered logographic. One, which we have termed the **distinct homophones** notion, is based on the idea that a word should be spelled in a way that is particular to that word, regardless of whether that results in different spellings for words that are homophonous. As we pointed out, this notion is what Sampson (1985) appealed to when he presented English orthography as a partly logographic system. The second was what we termed the **uniform spelling** notion, which prescribes that the same morpheme should be spelled the same way even if morphophonological changes result in different pronunciations for the morpheme in different environments. In this article, we have focused on the **distinct homophones** notion.

One obvious criticism of our approach is to suggest that we have merely redefined the term *logography* to mean a system in which homophonic words are spelled differently simply by virtue of their being different words, and that this goes against the spirit of what authors have meant by the term *logograph*. The most obvious reply to this criticism is that, as we showed in Section 2.1, previous authors are by no means

clear on precisely what is meant by the term "logography," and nothing in the way of a formal definition is ever given. It therefore seems reasonable to attempt to define the term by considering what procedure one might use to determine when one has a case of logography, and we have provided one such proposal in this article.

Still, one could insist that we have missed the point that in many systems that are considered logographic, there are components of written symbols that clearly do not indicate the pronunciation, but rather something to do with the meaning. To repeat an example we gave above: 琵琶 'Chinese lute' versus 枇杷 'loquat', both pronounced *pípá* in Mandarin, both having the same phonetic components 比巴 *biba*, and differing only in the semantic element used: 木 'tree' in the case of 'loquat' and 王王 for 'musical instrument' in the case of 'lute'. One could argue that these semantic elements are central to what authors have in mind when they use the term "logography." Of course, the methods we have presented in this article would certainly count these examples as logographic since the two words are homophonous, and one therefore requires context to determine which one is intended. But suppose for the sake of argument that only one of the words actually existed, say 枇杷 'loquat', and suppose also for the sake of argument that this was the only word pronounced *pípá*. Given that the written form still contains an obviously non-phonological element 木, would the fact that there were now no homophones make this example any less logographic?

We have two responses to this objection. The first is that while all originally invented writing systems of Mesopotamia, Egypt, China, and Meso-America share the characteristic of using a mix of semantic and phonological components, logography, as the term has been used in the literature, does not seem to consist merely in the existence of such elements in the writing system. Once again, Sampson (1985) argued that English spelling is at least partly logographic in that it makes a point of distinguishing in writing words that are homophonous in pronunciation. Clearly English has nothing equivalent to the 木 or 王王 semantic components, and so the presence of such components is not really a necessity for a system to count as logographic. In a similar vein the heterograms of Middle Persian, discussed in Section 2.2.2, which are widely considered as instances of logography, have nothing equivalent to Chinese semantic components.

Our second response is that to determine how logographic a system is, one has to consider the whole system, not just individual cases. While the removal of 琵琶 from the system would render 枇杷 non-logographic according to our measure, it would not have much of an effect on the logographicity of Chinese as a whole. To be sure, if there were never any homophonic words, then our measures would count the system as highly phonographic (S close to 0), no matter how many semantic components were still found as part of the written characters. But one would have to ask how likely such a system is to have occurred by chance: One would need a system in which there were no homophonic words at all (a priori unlikely) and in which nonetheless semantic elements were used to indicate some aspect of the meaning of the word. One could of course imagine something like such a system being developed by fiat: Suppose one started with a system like Chinese where there were in fact many homophonic words or morphemes, and where these were often distinguished in writing by additional components representing some aspect of the meaning of the word or morpheme in question. Suppose one then decided that every phonological unit (syllable or in some cases disyllable) should be represented by one and only one of the symbols used previously to distinguish among different words. The (originally) semantic components would still be present, though their semantic relevance would have been largely eliminated. Such a system does exist: the Yi syllabary (Shi 1996) mentioned above in the discussion of Penn and Choma (2006), which historically developed (by committee) from a Chinese-influenced

semantic-phonetic logographic script. Despite its logographic origins, the Modern Yi system is largely phonographic.[34]

We therefore believe that the criticisms cited above can be answered satisfactorily, and that the measures we have proposed for determining how logographic a system is are adequate for the one notion of logography we have addressed in this article. We furthermore believe that this notion of logography is sound. Of course, the measure is only as good as the corpora from which it is derived and the tools used to construct the corpora. We have noted the limitations of the tools in our discussion of the data preparation above. As for the corpora, obviously the Bible does not provide a full sample for any language, though we have also shown that the measures remain robust when we consider a different corpus, namely, Wikipedia. Still one might still expect the measures to be sensitive to the corpus used. Needless to say, this will be true of any attempt to formalize and quantify a notion such as logography: One is always dependent on the sample of the language that one chooses, and there is really no way to define the notion independent of specifying what data one is working with.

8. Conclusions

This article has explored an idea that was first introduced in Sproat (2000). Whereas writing systems can be classified into fairly distinct buckets according to what phonological units the systems represent, writing systems can also be classified according to how logographic they are. But this dimension is best treated as orthogonal to the phonological dimension, and as continuous rather than categorical.

We have argued that a previous attempt of Penn and Choma (2006) to quantify the difference between more or less logographic systems in fact fails to measure this distinction at all, but seems instead to be an artifact of using different text sizes in their two corpora.

We then presented an alternative approach, focusing on one notion of logography, which we have termed the **distinct homophones** notion. We have argued that one way to measure this is to consider the amount of attention paid outside the word in a simple attention-based RNN model when trying to spell that word given its pronunciation and the context in which it occurs, by an attention-based sequence-to-sequence model. This seems like an intuitively satisfying measure insofar as it relates to the intuition that in a highly logographic system, in order to know how to write a word, one must know

34 The above arguments justifying our notion of logography might still not satisfy those who would argue that writing systems like those of Chinese or Japanese, which make their logography overt in the form of the semantic radicals (Section 4), should be treated as a separate taxonomic class; two examples of such a position can be found in work of Joyce (2011) and Handel (2019). As noted in Section 1, Joyce argues against the term *logography*, favoring instead *morphography*, precisely because for him it is crucial that Chinese (and Japanese) characters represent *morphemes* rather than words, and of course as opposed to the letters of an alphabet, which basically represent sounds. The problem with this all-or-nothing view is that it has no way of describing systems like English orthography, other than to say that English spelling is just irregular. But this misses the point that English spelling irregularities function in much the same way as the more extreme and more overtly logographic systems of Chinese and Japanese, namely, to distinguish words or morphemes that would otherwise be written the same way. English orthography, the Perso-Aramaic heterograms we discussed in Section 2.2.2, and all other writing systems that have become more logographic over time are of course all secondary developments, unlike the case of Chinese, which has always had a strong logographic (or morphographic) component. Nevertheless, the fact that English uses a script where the basic elements are phonographic should not obscure the point that English orthography uses the elements of that script in ways that cannot be reconciled with the view that English is simply a phonographic writing system.

not only how it is pronounced, but which specific word or morpheme (among various possible identically pronounced words or morphemes) is involved.

This measure was compared to two computationally much simpler methods, namely, our *L* measures, which involved computing the mean number of spellings for given pronunciations found in a dictionary or corpus, as well as two *E* measures based on *n*-gram entropy. Although these also plausibly relate to the notion of logography, we have argued that they are ultimately less satisfactory than the attention measure *S* in the way they rank various writing systems. We also compared this measure to some more modern neural architectures, and while the results are comparable to what we achieved with the simpler model, we argued that they also show no great advantage over the simpler model.

In addition to the **distinct homophones** notion of logography, we also introduced the **uniform spelling** notion of logography, for which, however, we did not provide a method for measuring. We leave this problem to future work.

The results presented here are admittedly hard to evaluate, given that there is no defined evaluation set that ranks the degree of logography of various languages. One therefore cannot evaluate against "ground truth." We believe, however, that our results are useful for two reasons. First, we provide the first quantifiable measure of a notion that has been around for a long time, but has only occasionally been precisely defined. We noted in Section 2.1 that the **distinct homophones** notion of logography that we evaluate in this article corresponds exactly to the definition used by Handel (2019): Assuming Handel's definition is to be taken seriously, our work shows what that would mean for assessing how logographic a writing system is. Second, we believe that quantitative analysis of the kind we present here serves as a sort of baseline for future rigorous work on this topic.

Still, in principle we would like to correlate the results of our work reported here with results in the acquisition of reading and writing—with the acquisition of *spelling* being most directly relevant to the **distinct homophones** notion of logography. One would expect that the logographic rankings of writing systems reported here would correlate with the difficulty that native speakers of the language have in learning to spell in their writing system. Unfortunately, this is very hard to verify. Despite the existence of volumes that in principle deal with cross-linguistic comparison (e.g., Perfetti, Rieben, and Fayol 1997) there is really a dearth of careful cross-linguistic work. Part of the problem is that controlled comparison is hard, because reading and writing acquisition is obviously affected by the education system, when children learn to read and write, and how they are taught, and all these things vary substantially across countries. Another part of the problem is the overwhelming "Anglocentrism" of work on spelling and reading (Share 2008). And even when studies exist that compare spelling acquisition across languages, these invariably deal with just a handful of languages, often just two for a given study. That said, one study that does briefly survey prior work is Marinelli et al. (2015). Their own study compared spelling acquisition in primary school children in England and Italy, and found that Italian-speaking children were able to acquire accurate spelling after only two years of schooling, whereas English-speaking children still showed poor performance after 5 years. The paper also cites other studies that show similar results where a more "transparent" orthography is compared with a more "opaque" orthography, such as Czech versus English (Caravolas and Bruck 1993), Icelandic versus Danish (Juul and Sigurdsson 2005), and German versus English (Wimmer et al. 1991). While our experiments do not directly compare German and English, our Epitran experiments do compare German and French, and yield a substantially lower logography measure for German for S_{type}; whereas our main experiments show French

as being just slightly more logographic than English for S_{type}, suggesting that German would in any case count as less logographic than English, which is at least consistent with reported results for spelling acquisition. In general, the rankings in figures 10 and 14 for S_{type} are consistent with the finding in the literature on spelling acquisition that more opaque orthographies make it harder to learn to spell, the only surprise being Spanish, which we already discussed above.[35]

While our purpose in this article has been to investigate a question of interest in the study of writing systems, it may be reasonable to ask what broader applications results of this kind could have.[36] We do think that measures of the sort we propose could inform other areas. For example, assuming one can compute in the grapheme-to-phoneme direction a reasonable pronunciation for words in running text, our measure could highlight cases where transducing in the other direction, from phonemes to graphemes, could be potentially problematic. We noted a somewhat surprising instance of that, with Spanish, in Section 6.6. This could for example have implications for Automatic Speech Recognition, since in such cases the system has the potential to make a mistake in transcription; or in applied areas such as second-language learning (or even first language literacy), to identify potential areas where students may have problems.

Finally, we noted in the introduction that for the present ancient languages are outside the scope of the investigation since the methods we propose require a reasonable minimum amount of data. This is unfortunate insofar as all the originally invented ancient writing systems, and many of their derivatives had significant amounts of logography. We also mentioned in Section 2.2.2 the case of Middle Persian, which also had significant amounts of logography in the form of aramaeograms, Persian words that were spelled as the semantically equivalent Aramaic word. We would expect that the methods proposed in this article, applied to appropriately annotated corpora of Egyptian, Sumerian, Akkadian, Mayan, or Middle Persian would yield a high measure of logography for these writing systems. Hopefully, with further development of online corpora in such projects as the Cuneiform Digital Library Initiative,[37] Thesaurus Indogermanischer Text- und Sprachmaterialien (including Middle Persian),[38] or the Ramses Online project,[39] we may be able to address this problem in future work. If this becomes possible we would be able to quantify the degree to which the world's writing systems have become on balance less logographic over time, an interesting computational twist on Gelb's original intuition.

Appendix A: Further Details of Data Preparation

Table A.1 summarizes the details of data preparation (Section 6.1) for the languages used in the main experiments.

35 For a different approach to this issue see Beinborn, Zesch, and Gurevych (2016), who train a model to predict spelling difficulty, based on corpora of spelling errors in three languages.

36 We note in passing that such burden of proof of broader interest is inconsistently applied across areas of computational linguistics. For example, the authors of a paper on an improvement to machine translation or question answering would most likely not get such a question. On the other hand, the first author has in the past received comments on a paper on text normalization for speech synthesis that questioned the importance of the results since they seemed only to be of interest to TTS. Because one of the stated goals of the field of computational linguistics is to understand natural language phenomena via computational methods, such biases seem out of place.

37 `https://cdli.ucla.edu/`.

38 `https://titus.uni-frankfurt.de/indexe.htm?/texte/texte2.htm`.

39 `https://be.dariah.eu/project/ramses-online`.

Table A.1
Summary of the resources used for each of the languages.

Language	Phonetic Transcription	Additional packages/sources used
English	ARPAbet	https://pypi.org/project/pronouncing/
French	Idiosyncratic system	http://www.lexique.org/databases/Lexique383
Russian	Idiosyncratic system	https://github.com/kylebgorman/wikipron
Finnish	Finnish letters	
Swedish	SAMPA-derived	http://www.nb.no
Hebrew (Biblical)	Idiosyncratic system	https://www.mechon-mamre.org
Hebrew (Modern)	Idiosyncratic system	
Korean	Revised Romanization	https://pypi.org/project/ko-pron
Chinese	Pinyin	https://pypi.org/project/pinyin/
Japanese	Romaji	http://www.phontron.com/kytea https://github.com/chezou/Mykytea-python https://github.com/JRMeyer/jphones

Acknowledgments

We thank Kyle Gorman, Brian Roark, and Terry Joyce for helpful comments on an earlier version of this article. A version of this material was presented in Gorman's course on writing systems at the City University of New York, and at the ACL SIGTYP workshop in December 2020. We thank attendees for questions and comments. Finally, we thank three anonymous reviewers for *Computational Linguistics* for much critical feedback that helped us develop the ideas further.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, et al. 2016. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Aro, Mikko. 2017. Learning to read Finnish. In Ludo Verhoeven and Charles Perfetti, editors, *Learning to Read Across Languages and Writing Systems*. Cambridge University Press, Cambridge, pages 393–415.

Bahdanau, Dmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Machine translation by jointly learning to align and translate. In *Proceedings of 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Behnke, Maximiliana and Kenneth Heafield. 2020. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2664–2674, Online.

Beinborn, Lisa, Torsten Zesch, and Iryna Gurevych. 2016. Predicting the spelling difficulty of words for language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 73–83, Association for Computational Linguistics, San Diego, CA.

Berman, Ruth A. 1997. Modern Hebrew. Robert Hetzron, editor, *The Semitic Languages*. Routledge, London, pages 312–333.

Caravolas, Marketa and Maggie Bruck. 1993. The effect of oral and written language input on children's phonological awareness: A cross-linguistic study. *Journal of Experimental Child Psychology*, 55(1):1–30. https://doi.org/10.1006/jecp.1993.1001

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*, Harper and Row,

New York. https://doi.org/10.1007/s10579-014-9287-y

Christodoulopoulos, Christos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence.

Coulmas, Florian. 1989. *Writing Systems of the World*. Blackwell, Oxford. https://doi.org/10.1080/10888438.2016.1251437

Coulmas, Florian. 2003. *Writing Systems: An Introduction to Their Linguistic Analysis*. Cambridge University Press, Cambridge.

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2978–2988, Florence.

Daniels, Peter. 1996a. Methods of decipherment. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 141–159.

Daniels, Peter. 1996b. The study of writing systems. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 3–17.

Daniels, Peter. 2018. *An Exploration of Writing*. Equinox, Sheffield.

Dauphin, Yann N., Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941.

DeFrancis, John. 1989. *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press, Honolulu.

Dehaene, Stanislas. 2009. *Reading in the Brain*. Viking, London.

Dehghani, Mostafa, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2019. Universal transformers. In *7th International Conference on Learning Representations (ICLR)*, OpenReview.net, New Orleans, LA.

Deng, Yuntian, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, pages 1–13, Curran Associates, Inc., Montréal, Canada.

Deri, Aliya and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Association for Computational Linguistics, Berlin.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN.

Diringer, David. 1958. *The Alphabet: A Key to the History of Mankind*. Hutchinson's Scientific and Technical Publications, New York.

Doll, Chris. 2017. Korean rŏmanizatiŏn: Is it finally time for the library of congress to stop promoting Mccune-Reischauer and adopt the Revised Romanization scheme? *Journal of East Asian Libraries*, 2017(165):8.

Drucker, Johanna. 1995. *Alphabetic Labyrinth: The Letters in History and Imagination*. Thames & Hudson, London.

Ebden, Peter and Richard Sproat. 2014. The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3):1–21.

Elbro, Carsten. 2006. Literacy acquisition in Danish: A deep orthography in cross-linguistic light. In R. Malatesha Joshi and P. G. Aaron, editors, *Handbook of Orthography and Literacy*. Lawrence Erlbaum, Mahwah, NJ, pages 31–45.

Frith, Uta. 1985. Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall, and M. Coltheart, editors, *Surface Dyslexia: Cognitive and Neuropsychological Studies of Phonological Reading*. Erlbaum, Hillsdale, NJ, pages 301–330.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1243–1252, Sydney.

Gelb, Ignace. 1952. *A Study of Writing*. University of Chicago Press, Chicago.
Géron, Aurélien. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
Gershevitch, Ilya. 1979. The alloglottography of Old Persian. *Transactions of the Philological Society*, 77(1):114–190.
Glorot, Xavier and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256, Sardinia.
Gorman, Kyle, Gleb Mazovetskiy, and Vitaly Nikolaev. 2018. Improving homograph disambiguation with supervised machine learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1349–1352, Miyazaki.
Graff, David and Ke Chen. 2005. Chinese Gigaword. *LDC Catalog No.: LDC2003T09*, 1:58563–58230.
Gülçehre, Çaglar, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. 2019. Hyperbolic attention networks. In *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA.
Haile, Getatchew. 1996. Ethiopic writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 569–576.
Halpern, Jack, editor. 2013. *The Kodansha Kanji Learner's Dictionary: Revised and Expanded*. Kodansha, New York.
Handel, Zev. 2019. *Sinography: The Borrowing and Adaptation of the Chinese Script*. Number 1 in Language, Writing and Literary Culture in the Sinographic Cosmopolis. Brill, Leiden.
Harris, Roy. 1995. *Signs of Writing*. Routledge, London.
Hedlund, Turid, Ari Pirkola, and Kalervo Järvelin. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management*, 37(1):147–161.
Horn, Roger A. and Charles R. Johnson. 2012. *Matrix Analysis*. Cambridge University Press, Cambridge.
Hornkohl, Aaron D. 2019. Pre-modern Hebrew: Biblical Hebrew. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, 2nd edition, Routledge Language Family Series, Routledge, pages 533–570.
Horodeck, Richard. 1987. *The Role of Sound in Reading and Writing Kanji*. Ph.D. thesis, Cornell University, Ithaca, NY.
Istrin, Viktor A. 1965. *Vozniknovenie i razvitie pisma* / Возникновение и Развитие Письма, Soviet Academy of Sciences / Академия Наук СССР, Nauka / Наука, Moscow, USSR. In Russian.
Jenkins, John H., Richard Cook, and Ken Lunde. 2020. Unicode Han database (UniHan), Unicode Consortium. Standard Annex #38, Unicode Consortium. Revision 29.
Joyce, Terry. 2011. The significance of the morphographic principle for the classification of writing systems. *Written Language and Literacy*, 14(1):58–81.
Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing*, 2nd edition. Pearson.
Juul, Holger and Baldur Sigurdsson. 2005. Orthography as a handicap? A direct comparison of spelling acquisition in Danish and Icelandic. *Scandinavian Journal of Psychology*, 46(3):263–272. `https://doi.org/10.1111/j.1467-9450.2005.00456.x`
Kaiser, Lukasz, Aidan N. Gomez, and François Chollet. 2018. Depthwise separable convolutions for neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, OpenReview.net, Vancouver, BC, Canada.
Kalchbrenner, Nal, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alexander Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.
King, Ross. 1996. Korean writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 218–227.
Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
Kirchhoff, Katrin, Mark Hasegawa-Johnson, Preeth Jyothi, and Leanne Rolston. 2018. LanguageNet Grapheme-to-Phoneme Transducers. Statistical Speech Technology, University of Illinois. Online: `https://github.com/uiuc-sst/g2ps`. Accessed 17 August 2020.

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient Transformer. In *8th International Conference on Learning Representations (ICLR)*, Online.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver.

Kučera, Henry and W. Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press, Providence, RI.

Kudrinski, Maksim and Ilya Yakubovich. 2016. Sumerograms and akkadograms in Hittite: Ideograms, logograms, allograms or heterograms? *Altorientalische Forschungen*, 43(1-2):53–66.

Lee, Jackson L., Lucas F. E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 4223–4228, Marseille, France.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, Lisbon.

MacKay, David J. C. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

Magwa, Wiseman. 2002. The Shona writing system: An analysis of its problems and possible solutions. *Zambezia*, 29(1):1–11.

Mair, Victor. 1996. Modern Chinese writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 200–208.

Marinelli, Chiara, Christina Romani, Cristina Burani, and Pierluigi Zoccolotti. 2015. Spelling acquisition in English and Italian: A cross-linguistic study. *Frontiers in Psychology*, 6(1843).

Matsunaga, Sachiko. 1994. *The Linguistic and Psycholinguistic Nature of Kanji: Do Kanji Represent and Trigger only Meanings?* Ph.D. thesis, University of Hawaii, Honolulu.

Michel, Paul, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–11, Vancouver.

Milde, Benjamin, Christoph Schmidt, and Joachim Köhler. 2017. Multitask sequence-to-sequence models for grapheme-to-phoneme conversion. In *Proceedings Interspeech 2017*, pages 2536–2540, Stockholm, Sweden.

Mnih, Volodymyr, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 2204–2212, Montreal.

Moorehouse, Alfred. 1953. *The Triumph of the Alphabet*. Henry Schuman, New York.

Moradi, Pooya, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online.

Mortensen, David R., Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2710–2714, European Language Resources Association (ELRA), Paris.

Nasjonalbiblioteket. 2011. Leksikalsk Database for Svensk. `https://www.nb.no/sbfil/dok/nst_leksdat_se.pdf`. (In Norwegian). Online; accessed 17 August 2020.

Naveh, Joseph. 1982. *Early History of the Alphabet: An Introduction to West Semitic Epigraphy and Palaeography*. Magnes Press, The Hebrew University, Jerusalem.

Neubig, Graham and Shinsuke Mori. 2010. Word-based partial annotation for efficient corpus construction. In *The 7th International Conference on Language Resources and Evaluation (LREC)*, pages 2723–2727, Malta.

Neubig, Graham, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, OR.

New, Boris, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, &*

Computers, 36(3):516–524. https://doi.org/10.3758/BF03195598

Novak, Josef Robert, Nobuaki Minematsu, and Keikichi Hirose. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938. https://doi.org/10.1017/S1351324915000315

O'Connor, M. 1996. Epigraphic Semitic scripts. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 88–107.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Demonstrations*, pages 48–53, Minneapolis, MN.

Penn, Gerald and Travis Choma. 2006. Quantitative methods for classifying writing systems. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 117–120, New York.

Perfetti, Charles, Laurence Rieben, and Michel Fayol, editors. 1997. *Learning to Spell: Research, Theory, and Practice Across Languages*. Routledge, Mahwah, NJ.

Peters, Ben and André F. T. Martins. 2020. One-size-fits-all multilingual models. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 63–69, Association for Computational Linguistics, Online.

Pope, Maurice. 1975. *The Story of Decipherment: From Egyptian Hieroglyphs to Linear B*. Thames and Hudson, Hong Kong.

Pope, Maurice. 1999. *The Story of Decipherment: From Egyptian Hieroglyphs to Maya Script*, revised edition. Thames and Hudson, Hong Kong.

Raffel, Colin, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 2837–2846, Sydney.

Rao, Kanishka, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229, IEEE, South Brisbane.

Ravid, Dorit. 2005. Hebrew orthography and literacy. In R. Malatesha Joshi and P. G. Aaron, editors, *Handbook of Orthography and Literacy*. Routledge, London and New York, pages 339–363.

Ravishankar, Vinit, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online.

Roark, Brian, Cyril Allauzen, and Michael Riley. 2013. Smoothed marginal distribution constraints for language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–52, Association for Computational Linguistics, Sofia.

Roark, Brian, Michael Riley, Cyril Allauzen, Terry Tai, and Richard Sproat. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, System Demonstrations, pages 61–66, Jeju Island.

Robinson, Andrew. 2007. *The Story of Writing: Alphabets, Hieroglyphs & Pictographs*, 2nd edition. Thames & Hudson, London.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Rogers, Henry. 2005. *Writing Systems: A Linguistic Approach*. Blackwell, Malden, MA.

Rubio, Gonzalo. 2006. Writing in another tongue: Alloglottography in the Ancient Near East. In Seth L. Sanders, editor, *Margins of Writing, Origins of Cultures*, volume 2 of *Oriental Institute Seminars*. The Oriental Institute of the University of Chicago, Chicago, IL, pages 33–66.

Saeed, John. 1999. *Somali*, volume 10 of *London Oriental and African Language Library*. John Benjamins Publishing.

Salomon, Richard G. 1996. Brahmi and Kharoshthi. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 373–383.

Sampson, Geoffrey. 1985. *Writing Systems*. Stanford University Press, Stanford, CA.

Sampson, Geoffrey. 2012. *Writing Systems*, 2nd edition. Stanford University Press, Stanford, CA.

Sansom, George B. 1928. *An Historical Grammar of Japanese*. Oxford University Press, Oxford.

Schuster, Mike and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681. https://doi.org/10.1109/78.650093

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, Claude E. 1951. Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1):50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Share, David L. 2008. On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4):584. https://doi.org/10.1037/0033-2909.134.4.584

Shi, Dingxu. 1996. Yi script. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 239–243.

Silfverberg, Miikka, Francis Tyers, Garrett Nicolai, and Mans Hulden. 2021. Do RNN states encode abstract phonological processes? *arXiv preprint arXiv:2104.00789*.

Skjaervo, P. Oktor. 1996. Aramaic scripts for Iranian languages. In Peter Daniels and William Bright, editors, *The World's Writing Systems*, Oxford University Press, Oxford, pages 515–535.

Smith, Janet S. (Shibamoto). 1996. Japanese writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*, Oxford University Press, Oxford, pages 209–217.

Sproat, Richard. 2000. *A Computational Theory of Writing Systems*. ACL Studies in Natural Language Processing. Cambridge University Press, Cambridge.

Sproat, Richard. 2014. A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481. https://doi.org/10.1353/lan.2014.0042, https://doi.org/10.1353/lan.2014.0031

Sproat, Richard. 2016. English orthography among the writing systems of the world. In Vivian Cook and Des Ryan, editors, *The Routledge Handbook of the English Writing System*. Routledge, Milton Park, Abington.

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, pages 3104–3112, MIT Press.

Tang, Gongbo, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels.

Threatte, Leslie. 1996. The Greek alphabet. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, Oxford, pages 271–280.

van Daal, Victor H. P. and Malin Wass. 2017. First-and second-language learnability explained by orthographic depth and orthographic learning: A "natural" Scandinavian experiment. *Scientific Studies of Reading*, 21(1):46–59.

van de Velde, Mark, Koen Bostoen, Derek Nurse, and Gérard Philippson. 2019. *The Bantu Languages*, 2nd edition. Language Family Series, Routledge, London and New York.

van den Oord, Aäron, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, Long Beach, CA.

Vig, Jesse and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence.

Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne.

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multihead self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Association for Computational Linguistics, Florence.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, pages 5776–5788, Vancouver.

Wells, John C. 1997. SAMPA computer readable phonetic alphabet. In Dafydd Gibbon, Roger Moore, and Richard Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*, volume 4. Mouton de Gruyter, Berlin and New York, pages 684–732.

Wiegreffe, Sarah and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong.

Wimmer, Heinz, Karin Landerl, Renate Linortner, and Peter Hummer. 1991. The relationship of phonemic awareness to reading acquisition: More consequence than precondition but still important. *Cognition*, 40(3):219–249. https://doi.org/10.1016/0010-0277(91)90026-Z

Witten, Ian and Timothy Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094. https://doi.org/10.1109/18.87000

Woods, Christopher, Emily Teeter, and Geoff Emberling, editors. 2010. *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. Number 32 in Oriental Institute Museum Publications. Oriental Institute, Chicago.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, Lille.

Yolchuyeva, Sevinj, Géza Németh, and Bálint Gyires-Tóth. 2019. Transformer based grapheme-to-phoneme conversion. In *Proceedings of Interspeech 2019*, pages 2095–2099, Graz.

Yu, Fisher and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico.

Yu, Shiwen et al. 2002. The grammatical knowledge-base of contemporary Chinese – a complete specification (second version). Technical report, Tsinghua University Press, Beijing.

Zhang, Hao, Richard Sproat, Axel Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization. *Computational Linguistics*, 45:293–337. https://doi.org/10.1162/coli_a_00349

Zhao, Wei, Steffan Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing language-agnostic multilingual representations. *arXiv.org*, CoRR 2020.