# Comparative Analyses of the Information Content of Letters, Characters, and Inter-Word Spaces Across Writing Systems

**Linjieqiong Huang[1], Erik D. Reichle[2], & Xingshan Li[1,3]**

[1]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

[2]School of Psychological Sciences, Macquarie University, Sydney, Australia

[3]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

Author Note

This research was supported by two grants from the National Natural Science Foundation of China (32371156, 31970992). This work was also jointly funded by the National Natural Science Foundation of China and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 62061136001/DFG TRR-169. Linjieqiong Huang was supported by China Postdoctoral Science Foundation (2022M723362) and the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (E2CX6625CX).

Correspondence should be addressed to Xingshan Li, Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China. Email: lixs@psych.ac.cn.

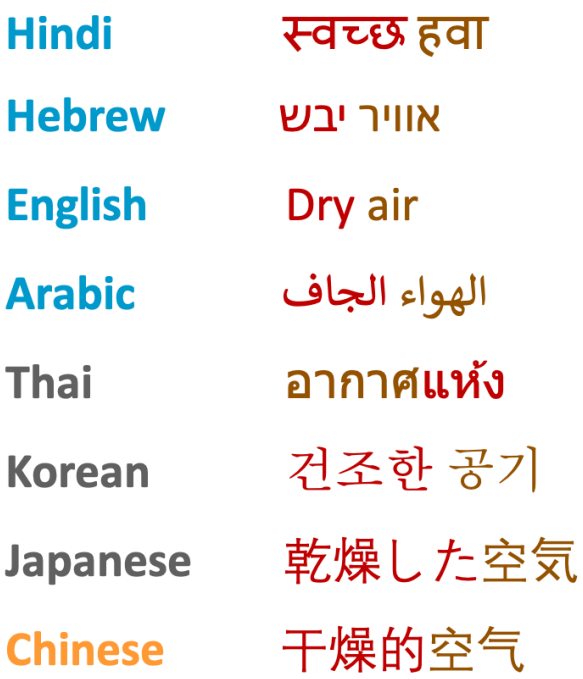
## Abstract

One difference among writing systems is how orthographic cues are used to demarcate words; whereas most alphabetic scripts use inter-word spaces, some Asian scripts do not explicitly mark word boundaries (e.g., Chinese). It is unclear whether these differences are arbitrary, or whether they are by design to maximize reading efficiency. Here we show that spaces inserted between words in non-demarcated scripts provide less information about word boundaries than spaces in demarcated scripts. Furthermore, despite the fact that less information is contained by inter-word spaces than characters/letters of the same size, the information content of inter-word spaces in demarcated scripts is closer to that of characters/letters compared to the information content of inter-word spaces that are inserted in non-demarcated scripts. These results suggest that the conventions used to demarcate word boundaries are sufficient to support efficient reading. Our findings provide new insights into the universals and variation across writing systems and shed light on the mental processes that support skilled reading.


**Keywords:** information theory, reading, scripts, writing systems

**Graphical Abstract**

The different methods for marking word boundaries that have been adopted by various writing systems are not arbitrary but are instead co-determined by other orthographic properties of the writing systems and have been adopted to increase reading efficiency.



## Introduction

Reading and writing are the most important technologies that humans have invented, allowing us to record history, literature, religious texts, science, and technology.[1] Because the archeological evidence indicates that this technology has only been available for five millennia, with literacy being commonplace for only a few centuries, reading and writing have not shaped human evolution but are painstakingly learned through years of formal education and practice. The end product of this is the capacity to coordinate the systems that support vision, attention, and spoken language to perform skills critical for success in modern, technologically advanced societies.[2]

As Table 1 shows, the writing systems or *scripts* used to record different languages vary widely. The majority languages use phonemically-based writing systems to convey the sounds of a language. While many of these writing systems incorporate both consonant and vowel information, some, like Arabic and Hebrew, exclusively represent consonants, omitting vowels. Arabic and Hebrew are also distinguished by their organized morphological systems, which rely on root letters and word patterns. A few other languages, like Chinese, use logographic characters to represent the meanings and sounds of morpho-syllables.[3] Another important difference and the focus of this article is whether scripts use orthographic cues to mark boundaries between words. As Figure 1 shows, most alphabetic scripts, like English, demarcate words using inter-word spaces. However, other scripts, like Hindi, employ additional cues to demarcate words. Written in the alpha-syllabic Devanagari script, Hindi uses ligatures or horizontal lines to connect letters belonging to the same word.[4,5] For the purpose of exposition, such scripts will be called “demarcated” throughout the remainder of this article. Such scripts remove any ambiguity about the locations of individual words, allowing their letters to be grouped into perceptual “objects” even when viewed in peripheral vision.

**Table 1 & Figure 1**

The demarcation of words, however, can be conceptualized as a continuum, with demarcated scripts on one end and "non-demarcated" scripts on the other. Again, Chinese provides an example. As Figure 1 shows, Chinese is written using continuous arrays of box-like characters, with individual words consisting of one or more characters without any type of explicit indicator of their boundaries or how the characters are grouped into their corresponding words.[6] The lack of explicit word boundaries can result in ambiguity. For example, the three-character string 花生长 can be segmented two different ways: with the first two characters as a word (花生-长 meaning "peanuts grow") or with the last two characters as a word (花-生长 meaning "flowers grow"). The lack of clear word boundaries also has implications for our understanding of eye-movement control in reading because the saccadic targets used in reading alphabetic scripts (centers of upcoming words) are less obvious in Chinese.[7]

Finally, as Figure 1 shows, there are several "partially" demarcated scripts that sometimes but not always provide cues about word boundaries. An example is Japanese, which is written without inter-word spaces using three types of characters: *kanji*, or characters borrowed from China for Sino-origin content words, *hiragana*, a phonetic syllabary mainly used for function words, and *katakana*, another phonetic syllabary mainly used for Western-origin loan words.[8] Moreover, because kanji is often visually denser than hiragana and katakana[8,9], transitions between kanji characters and the simpler hiragana and katakana (e.g., a kanji character surrounded by hiragana and/or katakana) can provide salient clues about word boundaries.

A second demarcated script is Thai, which also lacks inter-word spaces but is

alphabetic[10], with the locations of certain vowels providing cues about word boundaries. For example, vowels can be written above or below initial consonants (e.g., ปู meaning “crabs”, has a vowel ู below the consonant) or the second of two consonants (e.g., ไข่ meaning “eggs”, has a vowel ่ above the second consonant). When reading a phrase like ไข่และปู (“eggs and crabs”), the locations of the vowels (ู and ่) can be used to segment the words. However, in a phrase like “มะม่วงและทุเรียน” (“mango and durian”), where the words “mango” (มะม่วง) and “durian” (ทุเรียน) are polysyllabic, the vowels do not provide cues about word boundaries.

Finally, Korean is a third partially demarcated script. In Korean, spaces are not inserted between words but are inserted between *eojeol*, or parts of sentences comprised of one or more stem morphemes along with functional morphemes (e.g., case markers).[11] For instance, in the sentence 아버지가 방에 들어가신다 (“Father is going into the house.”), the particle 가 is used to indicate the subject. This particle is not separated from the preceding word 아버지 (“father”) because doing so would change the overall meaning of the sentence (e.g., 아버지 가방에 들어가신다 meaning “Father is going into the bag.”). Korean is therefore mid-way along the continuum of demarcated and non-demarcated scripts because spaces are used to separate stem morphemes but not stem and functional morphemes.[12]

These script-related differences have received relatively little attention by reading researchers. Because most reading research has focused on alphabetic languages like English, [3] the role of inter-word spaces during reading has been largely ignored except for acknowledging their likely role in guiding eye movements during reading, where word boundaries are presumably used to direct the eyes towards the centers of unidentified words in peripheral vision.[13] It has only recently been appreciated that inter-words spaces (or the lack thereof) play important roles in word identification.[14-17] It is therefore important to understand how developing readers of non-demarcated scripts learn to rapidly and accurately

segment arrays of letters/characters into sequences of meaningful words. This entails having a better understanding the roles played by different word-demarcation conventions across languages and scripts. Although the analyses reported in this article focus mainly on the contrast between Chinese (a non-demarcated script) and other demarcated and partially demarcated scripts, the conclusions that we draw from these analyses are quite general and have ramifications for our basic understanding of the mental processes that support skilled reading and how its development may be affected by differences among languages and writing systems.

Let us therefore begin by considering the consequences of using or not using inter-word spaces to mark word boundaries. The inclusion of inter-word spaces obviously makes text longer. Although this had practical ramifications historically, when writing materials were expensive, the main consequence today is that the inclusion of inter-word spaces reduces the perceptibility of text. Decades of research has demonstrated that the perception of letters and characters is limited to central vision.[18] As the distance between central vision and the letters/characters increases, the density of photoreceptors required to perceive them decreases while the lateral inference or "crowding" from spatially adjacent letters/characters increases, both of which reduce visual acuity. The presence of inter-word spaces thus degrades letter/character perception by pushing them further from central vision. This degradation might be offset, however, because letters/characters adjacent to spaces are subject to less crowding. This trade-off might also interact with other script attributes; for example, because Chinese characters are visually denser than letters, any cost from being further from central vision might outweigh any benefit from reduced crowding.

Another consequence of inter-word spaces is related to the cognitive effort required to segment lines of letters/characters into words. Whereas inter-word spaces allow readers to rapidly and accurately determine word boundaries from peripheral vision,[19,20] the absence of

spaces makes word segmentation effortful, necessitating a greater reliance on contextual and linguistic information. Although this reliance upon higher-level information might be viewed as being analogous to what happens with spoken language comprehension, it is worth noting that the latter is also reliant upon various sub-lexical cues (e.g., phonotactics, acrostic-phonetics, word stress[21]) that, like the inter-word spaces in scripts, can be used to aid word segmentation. The absence of inter-word spaces therefore necessitates the use of contextual and linguistic knowledge to aid in word segmentation, thereby requiring additional cognitive effort. (It is also noteworthy that, even with adequate contextual and linguistic information, word segmentation is not always unambiguous or correct.[22]) Thus, with everything being equal, inter-word spaces are predicted to increase reading efficiency by decreasing the cognitive effort required to segment and identify words. But what exactly is meant by "cognitive effort" in this context?

One suggestion is that cognitive effort can be quantified using metrics of information availability.[23,24] Several experiments support this conjecture by showing that informative words require more time to read.[25,26] Although these experiments have estimated the information content of the words themselves, the information provided by inter-word spaces might also contribute to the effort required to segment/identify words. This hypothesis might explain the lack of consistency in the use of inter-word spaces across scripts: If spaces provide minimal information about word boundaries, then their omission might produce negligible word-segmentation/identification cost, while their inclusion might reduce the perceptibility of words by pushing them further from central vision. To the best of our knowledge, this possible trade-off has not been tested.

To test the possibility that different scripts adopt different word-demarcation conventions to support efficient reading given the constraints imposed by their other orthographic features, this study investigated whether inter-word spaces in demarcated scripts

contain more information about word boundaries than spaces inserted between words in both non-demarcated and partially demarcated scripts. To do this, the information content of inter-word spaces was estimated in 27 different scripts (Table 1). These scripts are widely used, representative of many world languages, and include examples of demarcated, partially, and non-demarcated scripts. We measured how informative inter-word spaces are for determining word boundaries for demarcated scripts, and estimated how informative inter-word spaces are if they are inserted into partially and non-demarcated scripts.

To quantify informativeness, we utilized *information theory*, a mathematical framework used for quantifying, storing, and communicating information.[27-29] According to information theory, if an event's occurrence is highly uncertain, then it has more information content. And conversely, if the occurrence of an event is highly predictable, then it has less information content. As described by Equation 1, *entropy*, $H$, is a measure of this uncertainty using the probability distribution of some event represented by a random variable, $x$; the higher the entropy, the greater the uncertainty of the random variable, indicating more information content.[29] The information content of an event is defined as the negative logarithm of its probability of occurrence, $p$.[24] By summing the information content of all possible events and taking their average, the value of entropy is obtained. When using base-2 logarithms, the unit of measurement for entropy is *bits*. Therefore, entropy measured in bits provides a means to quantify the amount of information content.

$$H = \sum p(x) \log \frac{1}{p(x)} \quad (1)$$

There has been a growing body of research delving into the cognitive mechanisms of language processing through the lens of information theory.[28,30-32] From the perspective of information theory, it has been proposed that word length is primarily determined by the

average information content contained by a word within a given context.[24,33,34] But the informativeness of a word also depends on its context. Consequently, to maintain a relatively constant number of communicated bits per unit of time, if a word is highly predictable within a given context and the demand for information is low, then a shorter form can be used to designate the word. Conversely, a word that has low predictability contains more information content, necessitating a longer form to designate the word. As these examples illustrate, information theory is a valuable tool and framework for understanding such trade-offs and the workings of languages more generally, although the rules for demarcating words, the ocus of this study, has remained an unexplored issue.

With the objective of better understanding the aforementioned trade-offs, we applied Equation 2 to quantify the reduction in the uncertainty about the locations of word boundaries when adding inter-word spaces, in alignment with information theory. Inter-word spaces mark word boundaries by indicating the terminal position of a word. Therefore, the amount of information about word boundaries contained by inter-word spaces can be understood as the amount of information gained by knowing the letter/character that terminates a word. Because the length of a word reflects the position of its final letter/character, we used the proportion of words of different lengths to calculate information content, which is equivalent to the probability of the $i^{th}$ letter/character terminating a word. As Equation 2 indicates, $n$ denotes the maximal word length in a given language and $p_i$ denotes the probabilities of words with $i$ letters/characters in natural connected discourse.

$$H = \sum_{i=1}^{n} p_i log_2 \frac{1}{p_i} \quad (2)$$

Furthermore, because an inter-word space usually occupies a letter/character-sized space in fully-demarcated written languages, we compared the information content of inter-

word spaces with the average information content of letters/characters in the same script to further evaluate the efficiency of inter-word spaces. If inter-word spaces contain approximately the same amount of information as actual letters/characters, then they might be more likely to be adopted to demarcate words. Otherwise, it might be more economical to not use inter-word spaces. Regarding the information content of letters/characters, we used Equation 2 with different values of $p_i$, representing the probability of occurrence of letter/character $i$ in a large corpus. The information content calculated in this way corresponds to the reduction of uncertainty that comes from seeing a letter/character, under the assumption that letters/characters are mutually independent.[35] We should note that although this assumption is technically not true because letters/characters and words respectively occur in the contexts of words and larger passages of text, we proceeded as indicated because we were primarily interested in comparing the information content of inter-word spaces with that of letters/characters.

The trade-off between the having condensed texts and thus enhanced perceptibility, on one hand, and the reduction of cognitive processing effort that comes from explicitly marking word boundaries, on the other, may be the reason why different writing systems adopt different word-demarcation conventions. This leads to an obvious prediction: If the conventions used by a script to demarcate word boundaries support efficient reading given its other orthographic constraints, then the inter-word spaces in demarcated scripts will contain more information than if spaces are inserted between words in both non- and partially demarcated scripts. Saying this another way, the informativeness of inter-word spaces should be more comparable to that of letters/characters in demarcated scripts than in non- and partially demarcated scripts.

## Materials and Method

Using Equation 2, we first quantified both the information content of inter-word

spaces in demarcated scripts and the information content of inter-word spaces when inserted in non- and partially demarcated scripts. To calculate $p_i$, we calculated the number of occurrences of all words of a specific length, and then divided this quantity by the total number of words in the corpus. To capture the influence of natural connected discourse, the word-length distributions were based on token frequencies. Because word-length distributions differ across languages, the values of $p_i$ also vary (see Figure 2). The word-length distributions for 26 languages except Chinese were calculated using word frequency lists from OpenSubtitle tokenized source.[36,37] The OpenSubtitle corpus is derived from an extensive database of movie and TV subtitles, encompassing 1,689 bitexts extracted from subtitle files, totaling 2.6 billion sentences (17.2 billion tokens) in over 60 languages. In the case of Chinese, the word frequency list was obtained from the SUBTLEX-CH, a database of word frequencies derived from a corpus of films and TV subtitles, totaling 46.8 million characters and 33.5 million words.[38] Although these different corpora may not be comparable for some purposes (e.g., examining the prevalence of different syntactic structures), all of the corpora are large and thus likely to provide representative estimates of the types of information required for our analyses (e.g., statistics about the lengths and frequencies of words).

**Figure 2**

To compare the information content of inter-word spaces versus letters/characters, we quantified the average information content of letters/characters within each script and then computed the ratio of information contained by inter-word spaces to that of letters/characters. We again used Equation 2, but used values of $p_i$ corresponding to the probability of occurrence of letter/character $i$ in a large corpus. To calculate $p_i$, we first obtained all the

letters/characters and calculated the number of occurrences of each letter/character based on the aforementioned word frequency list for each language. We then divided this value by the total number of occurrences of all letters/characters in the entire list. The calculations of the information content contained by inter-word spaces and letters/characters are thus based on homogeneous corpora, ensuring the comparability of the results.

## Results

The information content of inter-word spaces showed different patterns for scripts using different conventions for indicating word boundaries (Figure 3). First, the inclusion of spaces provided less information about word boundaries in non-demarcated scripts like Chinese (1.10 bits) than in partially demarcated scripts ($M$ = 2.23 bits, $SD$ = 0.73). Second, the inclusion of spaces in partially demarcated scripts provided less information about word boundaries than spaces in demarcated scripts ($M$ = 3.01 bits, $SD$ = 0.23). Finally, spaces added between words in Thai provided more information (3.21 bits) compared with other partially demarcated scripts ($M$ = 1.73 bits, $SD$ = 0.26).

**Figure 3**

These results might reflect differences in the variations of word lengths across languages[1]. As shown in Figure 2, the word-length distributions varied considerably. In Chinese, where most words do not exceed four characters, the average word length was shorter and less variable ($M$ = 1.40 characters, $SD$ = 0.57) compared to both demarcated ($M$ = 4.33 letters, $SD$ = 2.35) and partially demarcated ($M$ = 2.41 characters/letters, $SD$ = 1.49)

[1] In quantifying word length across languages, we used the LEN function in Microsoft Excel. Specifically, we measured word length using characters in Chinese and using letters in alphabetic languages. For languages where the letters are connected by ligatures (e.g., Arabic), word length was also measured using letters. For Korean, we measured word length using syllables because their boundaries are consistent (as per character boundaries in Chinese).

scripts. For that reason, there is less uncertainty about the lengths of, and hence the likely boundaries between, Chinese words. Consequently, inserting inter-word spaces between Chinese words provides less additional information about word boundaries compared to both demarcated and partially demarcated languages.

The comparison of the information content of inter-word spaces versus letters/characters also showed different patterns for scripts utilizing different word-demarcation conventions (Figure 3). First, characters in Chinese, a non-demarcated script, contain more information (9.84 bits) than letters/characters in both demarcated ($M$ = 4.47 bits, $SD$ = 0.28) and partially demarcated ($M$ = 7.75 bits, $SD$ = 1.60) scripts. The reason for these differences is that (as shown in Table 1) there are more characters in Chinese than letters/characters in demarcated and partially demarcated scripts, with the identification of a letter/character comes from a larger set by definition resulting in a larger reduction in uncertainty. The identification of a Chinese character thus reduces uncertainty more than the identification of a letter/character in other scripts. This suggests that Chinese characters carry more information content than their counterparts in other scripts.

Second, in Chinese, the ratio of information contained by inter-word spaces versus characters is 0.11, indicating that spaces contain much less information than characters of the same size. In contrast, in demarcated scripts, the same ratio is 0.68 ($SD$ = 0.07), indicating that spaces contain a more comparable amount of information to letters/characters of the same size. And in partially demarcated scripts, the ratio is in between ($M$ = 0.32, $SD$ = 0.18).

The preceding results collectively suggest that the insertion of spaces between Chinese words may not be economical because it pushes the characters away from the central vision, reducing their perceptibility and thereby reducing the information content. Nevertheless, it is noteworthy that, in all instances of demarcated, partially-, or non-demarcated scripts, inter-word spaces contain less information than letters/characters.

Although this observation might suggest that it would be beneficial to remove inter-word spaces to enhance the perceptibility of text, we earlier reviewed evidence that the elimination of inter-word spaces can decrease reading efficiency by making the segmentation and identification of words more difficult and/or prone to error.

## Discussion

Writing systems differ in their use of explicit word-boundary markers. Our research suggests that these differences are not arbitrary but instead reflect the specific demands imposed by different scripts. For example, inter-word spaces provide more information in demarcated than non-demarcated scripts. This difference may reflect the fact that demarcated scripts are inherently subject to more ambiguity regarding word boundaries if spaces are removed, which would be expected to increase the cognitive effort required to segment words. In addition, although the information contained by inter-word spaces is less than that of characters/letters of the same size, the information content of inter-word spaces in demarcated scripts is closer to that of characters/letters compared to the information content of inter-word spaces in non-demarcated scripts. Consequently, any benefit that may be derived from inserting spaces in non-demarcated scripts may be offset because the spaces push the characters away from the central vision, reducing their perceptibility. However, in demarcated scripts, the reduced perceptibility of letters/characters that stems from the insertion of inter-word spaces can be partially offset by the information provided by the spaces. Whether and how a script marks word boundaries thus reflects a trade-off between making texts more compact and thus more perceptible, on one hand, and reducing the cognitive effort required to segment words, on the other. Different scripts adopt different word-demarcation conventions to support efficient reading given the constraints imposed by their other orthographic features.

The current investigation has also shown that the methods used to indicate word

boundaries in partially demarcated scripts may also be close to optimal. Consistent with this conjecture, our findings suggest that the insertion of spaces in Japanese and Korean provides less information about word boundaries than the spaces in demarcated scripts, but more than inserted spaces in non-demarcated scripts. Additionally, the ratio of the information contained by inter-word spaces versus characters/letters in Japanese and Korean is between that of demarcated and non-demarcated scripts. And with Thai, although the information contained by inter-word spaces is comparable to that in demarcated scripts, the ratio of information contained by spaces versus letters is greater than in non-demarcated scripts but less than in most demarcated scripts. These results suggest why it may be economical to separate only parts of sentences (e.g., eojeol in Korean) rather than individual words in partially demarcated scripts.

The observed differences in the word-length distributions and the numbers of different letters/characters might also contribute to differences in the informativeness of inter-word spaces across languages. As previously mentioned, according to information theory, the uncertainty of an event contributes to its information content. As applied to Chinese, where words are shorter and exhibit less variability in length, the uncertainty regarding word length, which can function as a proxy for knowing a word's boundaries, is reduced compared to both demarcated and partially demarcated scripts. Furthermore, the number of letters/characters types naturally causes variation in word length. To optimize communication, it is more efficient to use words of different lengths, using short forms for frequent words and long forms for infrequent words.[39] Because there are more characters in Chinese than there are letters in alphabetic scripts, each character contains more information than the letters in demarcated and partially demarcated scripts. For that reason, most Chinese words can be represented by one or two characters, whereas most words in alphabetic scripts require at least a few letters. This suggests that the inclusion of inter-word spaces may therefore also be

influenced by how a language opts to represent phonology.

Our conclusions are broadly consistent with the results of studies that have examined the consequences of removing spaces in demarcated scripts or adding spaces in partially or non-demarcated scripts (see Figure 4). For example, in English, removing spaces reduced reading rate by almost 50%.[20,40] Conversely, in Chinese, inserting inter-word spaces did not affect reading speed.[41] And in partially demarcated scripts, adding inter-word spaces did not significantly improve reading speed.[8,10,42-45] These results are consistent with the utility of inter-word spaces being determined by the cognitive effort required to segment and identify words. In demarcated scripts, inter-word spaces reduce the effort required to segment/identify words, so that the removal of spaces hinders reading. However, in scripts where inter-word spaces are less informative, their insertion does not facilitate reading.[2] Although these results are consistent with our interpretation of the present study, it is important to acknowledge another possible explanation.

This alternative is related to the influence of reading-format familiarity. In demarcated scripts, where readers are accustomed to a spaced format, the removal of inter-word spaces might be unfamiliar and thus problematic. By contrast, in Chinese, where readers are familiar with an unspaced format, readers may have developed an efficient word-segmentation process that results in little or no added benefit when spaces are inserted between words[41]. And in the case of partially demarcated scripts such as Japanese, readers may be accustomed to using other cues (e.g., the visual distinctiveness of characters) to help segment words. If this alternative account is correct, then it suggests that format familiarity plays a significant role in reading efficiency. It also suggests that further studies are needed to distinguish between these two explanations.

[2] The potential benefit of adding spaces might be offset because the words are further from central vision and thus less perceptible, and/or because non-demarcated scripts are unfamiliar.

**Figure 4**

Although adult readers in partially and non-demarcated scripts do not benefit from the insertion of spaces between words, developing readers (e.g., children and second language learners) certainly do. In the case of Chinese, the introduction of inter-word spaces has been shown to help children form stronger connections between characters and words, thereby supporting the more rapid and accurate learning of new vocabulary.[46] There is also evidence suggesting that explicit word boundary information provided by alternating the font color of words can enhance both silent and oral reading for Chinese beginners.[47-49] And in the case of second language learners, the insertion of spaces between words reduces the uncertainty about the characters that constitute a word, thereby improving word identification and reading speed.[50] Finally, it is worth noting that, with the partially demarcated Thai script, children start learning to read text with inter-word spaces in kindergarten and first grade, only transitioning to text without spaces in second grade. And similarly, with the partially demarcated Japanese script, children are initially taught to read spaced hiragana text.[8,10] The available evidence thus suggests that, for both non- and partially demarcated scripts, the insertion of spaces between words is beneficial for facilitating the learning process in developing readers.

Our study also contributes to understanding how written systems may have developed, particularly with respect to their current use of inter-word spaces. A historical analysis of alphabetic scripts indicates that the conventions used to mark word boundaries have changed over time. Historically, alphabetic scripts did not mark word boundaries,[51] either because the primary goal was to transcribe spoken language or because the writing materials were expensive. To comprehend the meaning of written texts, readers had to read the texts orally, despite this being less efficient than silent reading. This changed in the 7th

and 8th centuries, when Irish and Anglo-Saxon scribes introduced inter-word spaces, which then became standard in Renaissance Italy, France, and Byzantium at the end of the 16th century and in Slavic countries in the 17th century.[51] The historical evidence suggests that marking word boundaries was associated with increasing mass literacy and the need to increase reading efficiency.

Given these historic trends, one outstanding question is whether current writing systems could be improved to support even more efficient reading? Information theory might be useful for answering this question and guiding future writing-system reforms. For example, the Korean writing system uses spaces to demarcate parts of sentences but not words, allowing efficient reading of Korean. This convention might therefore improve the reading efficiency of scripts having similar information content of spaces versus letters/characters.

It is also important to note that a better understanding of script-related differences may also advance our understanding of the universal versus script-specific mental processes that support skilled reading. For example, because what constitutes a "word" is not constrained by clear boundaries in Chinese, it may afford more flexible lexical processing, allowing skilled readers to represent larger "chunks" of characters as "words" than less skilled readers. This conjecture is consistent with the lack of consensus among Chinese readers regarding the precise locations of word boundaries.[52] By way of comparison, demarcated scripts have fixed word boundaries, resulting in more invariant but stable word representations. Future research is required to test this hypothesis.

## Conclusion

The question of how individual words are separated stands out as an underappreciated but crucial distinction among written languages. We posit that each script has adopted a word-demarcation convention that suits its particular needs, and that if and how

word boundaries are denoted are not arbitrary, but instead support efficient reading given the other orthographic constraints of a script. Our findings contribute to a better understanding of both the divergence and consistency of reading across writing systems, which is critical for having a better understanding the universal and script-specific mental processes that support skilled reading.[3] Our findings are also of practical significance, offering possible insight into the reformation of writing systems and shedding light on the relationship between writing systems and cognition during reading.

## Acknowledgments

This research was supported by two grants from the National Natural Science Foundation of China (32371156, 31970992). This work was also jointly funded by the National Natural Science Foundation of China and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 62061136001/DFG TRR-169. Linjieqiong Huang was supported by China Postdoctoral Science Foundation (2022M723362) and the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (E2CX6625CX).

## Competing Interest Statement

The authors have no conflicts of interest to declare that are relevant to the content of this article.

## Authors' Contributions

Linjieqiong Huang: Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing-Original draft preparation, Writing-Reviewing and Editing.

Erik D. Reichle: Writing-Reviewing and Editing.

Xingshan Li: Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Supervision, Funding acquisition, Writing-Reviewing and Editing.

## Data Availability Statement

Data and code associated with this study are freely available in a public repository at: https://doi.org/10.57760/sciencedb.psych.00110.

## References


1. Sagan, C. (1980). *Cosmos*. London, UK: Abacus.
2. Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*, 31–74. https://doi.org/10.1111/1529-1006.00004
3. Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology, 1*(3), 133–144. https://doi.org/10.1038/s44159-022-00022-6
4. Jain, D., & Cardona, G. (2007). *The Indo-Aryan Languages*. Routledge.
5. Rao, C., Vaid, J., Srinivasan, N., & Chen, H. C. (2011). Orthographic characteristics speed Hindi word naming but slow Urdu naming: Evidence from Hindi/Urdu biliterates. *Reading and Writing*, *24*, 679–695. https://doi.org/10.1007/s11145-010-9256-9
6. Yu, L. & Reichle, E. D. (2017). Chinese vs. English: Insights on cognition during reading. *Trends in Cognitive Sciences*, *21*, 721–724. https://doi.org/10.1016/j.tics.2017.06.004
7. Wei, W., Li, X., & Pollatsek, A. (2013). Word properties of a fixated region affect outgoing saccade length. *Vision Research*, *80*, 1–6.
8. Sainio, M., Hyona, J., Bingushi, K., & Bertram, R. (2007). The role of interword spacing in reading Japanese: An eye movement study. *Vision Research, 47*(20), 2575–2584. https://doi.org/10.1016/j.visres.2007.05.017
9. Kajii, N., Nazir, T. A., & Osaka, N. (2001). Eye movement control in reading unspaced text: The case of the Japanese script. *Vision Research*, *41*(19), 2503–2510. https://doi.org/10.1016/S0042-6989(01)00132-8
10. Kasisopa, B., Reilly, G. R., Luksaneeyanawin, S., & Burnham, D. (2013). Eye

movements while reading an unspaced writing system: The case of Thai. *Vision Research, 86*, 71–80. https://doi.org/10.1016/j.visres.2013.04.007

11. Baek, H., Choi, W., & Gordon, P. C. (2023). Reading spaced and unspaced Korean text: Evidence from eye-tracking during reading. *Quarterly Journal of Experimental Psychology*, *76*(5), 1072–1085. https://doi.org/10.1177/17470218221104
12. Song, J. J. (2006). *The Korean language: Structure, use and context*. Routledge.
13. McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations in words. *Vision Research, 28*, 1107–1118.
14. Li, X., & Pollatsek, A. (2020). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review, 127*(6), 1139–1162. https://doi.org/10.1037/rev0000248
15. Li, X., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive Psychology, 58*(4), 525–552. https://doi.org/10.1016/j.cogpsych.2009.02.003
16. Ma, G., Li, X., & Rayner, K. (2014). Word segmentation of overlapping ambiguous strings during Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1046–1059. https://doi.org/10.1037/a0035389
17. Yu, L., Liu, Y., & Reichle, E. D. (2021). A corpus-based vs. experimental examination of word- and character-frequency effects in Chinese reading: Theoretical implications for models of reading. *Journal of Experimental Psychology: General*, *150*, 1612–1641. https://doi.org/10.1037/xge0001014
18. Veldre, A., Reichle, E. D., Yu, L., & Andrews, S. (2023). Understanding the visual constraints on lexical processing: New empirical and simulation results. *Journal of Experimental Psychology: General*, *152*, 693–722. https://doi.org/10.1037/xge0001295

19. Perea, M., & Acha, J. (2009). Space information is important for reading. *Vision Research, 49*(15), 1994–2000. https://doi.org/10.1016/j.visres.2009.05.009

20. Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision Research, 38*(8), 1129–1144. https://doi.org/10.1016/s0042-6989(97)00274-5

21. Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*(4), 477–500. https://doi.org/10.1037/0096-3445.134.4.477

22. Huang, L., Staub, A., & Li, X. (2021). Prior context influences lexical competition when segmenting Chinese overlapping ambiguous strings. *Journal of Memory and Language*, *118*, 104218. https://doi.org/10.1016/j.jml.2021.104218

23. Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science, 30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

24. Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

25. Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science, 42*, 1166–1183. https://doi.org/10.1111/cogs.12597

26. Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

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

28. Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

29. Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.
30. Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004
31. Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. https://doi.org/10.1126/science.1218811
32. Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science*, *10*(1), 209–224. https://doi.org/10.1111/tops.12316
33. Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108
34. Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010
35. Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x
36. Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 923–929). European Language Resources Association.
37. Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Lrec* (Vol. 2012, pp. 2214–2218). Citeseer.

38. Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS one*, *5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

39. Zipf, G. (1949). *Human behavior and the principle of least effort.* Addison-Wesley, New York.

40. Winskel, H., Radach, R., & Luksaneeyanawin, S. (2009). Eye movements when reading spaced and unspaced Thai and English: A comparison of Thai–English bilinguals and English monolinguals. *Journal of Memory and Language, 61*(3), 339–351. https://doi.org/10.1016/j.jml.2009.07.002

41. Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1277–1287. https://doi.org/10.1037/0096-1523.34.5.1277

42. Baek, H., Choi, W., & Gordon, P. C. (2022). Reading spaced and unspaced Korean text: Evidence from eye-tracking during reading. *Quarterly Journal of Experimental Psychology, 76*(5), 1072–1085. https://doi.org/10.1177/17470218221104736

43. Kohsom, C., & Gobet, F. (1997). Adding spaces to Thai and English: Effects on reading. *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, p. 388–393. Hillsdale, NJ: Erlbaum.

44. Leung, T., Boush, F., Chen, Q., & Al Kaabi, M. (2021). Eye movements when reading spaced and unspaced texts in Arabic. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43. Retrieved from https://escholarship.org/uc/item/56b348fk

45. Winskel, H., Perea, M., & Ratitamkul, T. (2012). On the flexibility of letter position coding during lexical processing: Evidence from eye movements when reading Thai. *Quarterly Journal of Experimental Psychology, 65*(8), 1522–1536.

https://doi.org/10.1080/17470218.2012.658409

46. Blythe, H. I., Liang, F., Zang, C., Wang, J., Yan, G., Bai, X., & Liversedge, S. P. (2012). Inserting spaces into Chinese text helps readers to learn new words: An eye movement study. *Journal of Memory and Language*, *67*(2), 241–254. https://doi.org/10.1016/j.jml.2012.05.004
47. Perea, M., & Wang, X. (2017). Do alternating-color words facilitate reading aloud text in Chinese? Evidence with developing and adult readers. *Memory & Cognition*, *45*, 1160–1170. https://doi.org/10.3758/s13421-017-0717-0
48. Pan, J., Liu, M., Li, H., & Yan, M. (2021). Chinese children benefit from alternating-color words in sentence reading. *Reading and Writing*, *34*, 355–369. https://doi.org/10.1007/s11145-020-10067-9
49. Song, Z., Liang, X., Wang, Y., & Yan, G. (2021). Effect of alternating-color words on oral reading in grades 2–5 Chinese children: Evidence from eye movements. *Reading and Writing*, *34*(10), 2627–2643. https://doi.org/10.1007/s11145-021-10164-3
50. Shen, D., Liversedge, S. P., Tian, J., Zang, C., Cui, L., Bai, X., Yan, G., & Rayner, K. (2012). Eye movements of second language learners when reading spaced and unspaced Chinese text. *Journal of Experimental Psychology: Applied*, *18*(2), 192–202. https://doi.org/10.1037/a0027485
51. Saenger, P. (1997). *Space between words: The origins of silent reading*. Stanford University Press.
52. Liu, P., Li, W., Lin, N., & Li , X. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading? *PLoS ONE, 8*(2), e55440. https://doi.org/10.1371/journal.pone.0055440
53. Chang, L.-Y., Plaut, D. C., & Perfetti, C. A. (2015). Visual complexity in orthographic learning: Modeling learning across writing system variations. *Scientific Studies of*

*Reading, 20*(1), 64–85. https://doi.org/10.1080/10888438.2015.1104688

**Figure 1. Different word-demarcation methods.**

Chinese is a non-demarcated script with no explicit markers for word boundaries. Thai, Korean, and Japanese are partially demarcated scripts because word boundaries are unreliably indicated. Hindi, Hebrew, English, and Arabic are demarcated scripts because word boundaries are unambiguously indicated. All examples represent the phrase "dry air," with the words in each language being respectively represented in red and brown font.

**Figure 2. Word-length distribution of 27 written languages.**

The bar graphs show the distributions of word lengths in different languages. Information content contained by inter-word spaces was calculated using word-length distributions based on token frequencies (green bars). For comparison, word-length distributions based on type frequencies (pink bars) are also presented.

**Figure 3. Information content contained by inter-word spaces and characters/letters.**

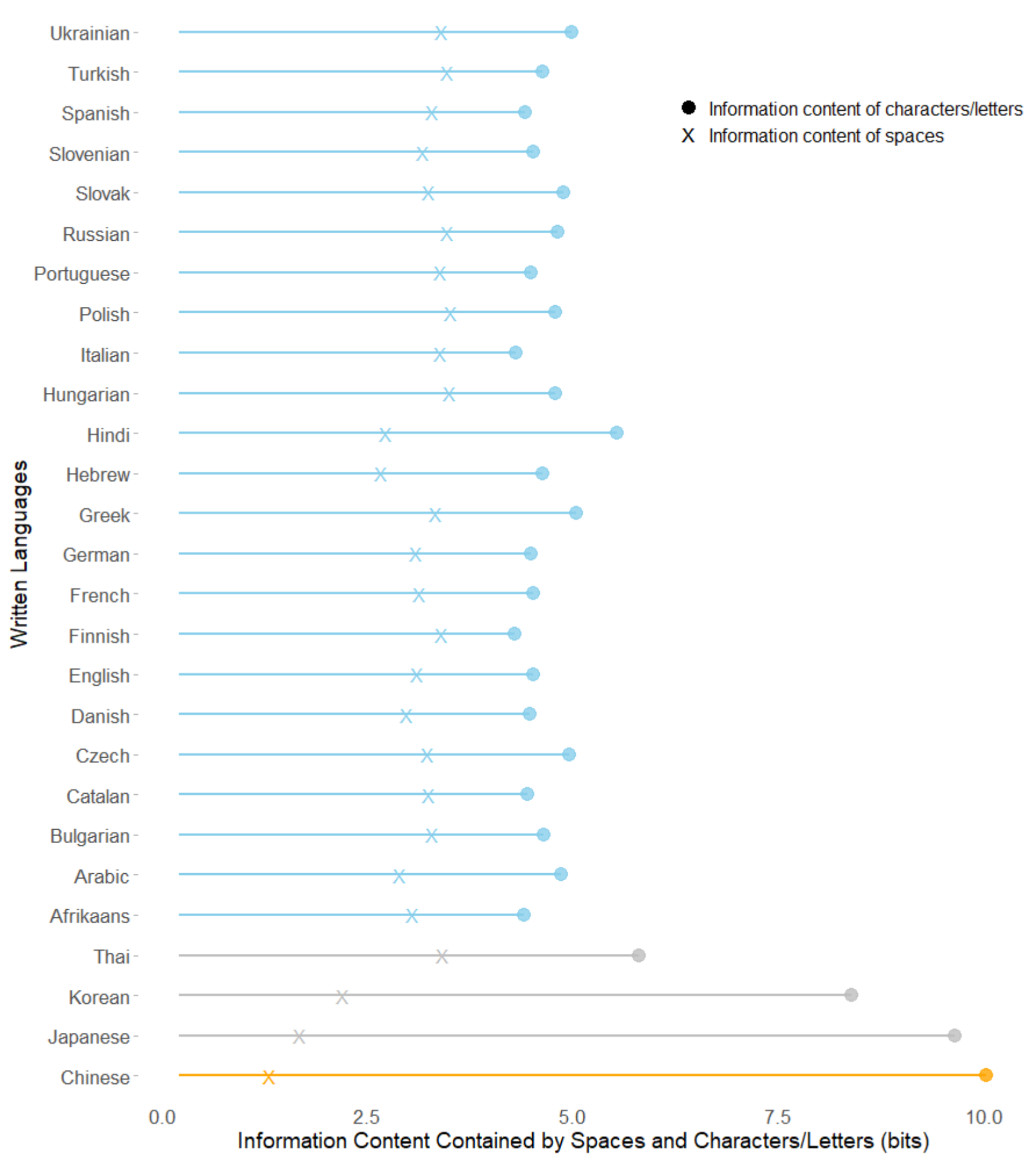
The information content contained by letters/characters and inter-word spaces is indicated by "●" and "×," respectively. Blue lines correspond to demarcated scripts, grey lines correspond to partially demarcated scripts, and the orange line corresponds to Chinese, a non-demarcated script.

**Figure 4. Reading rates with vs. without inter-word spaces.**

Reading rates are measured in characters per minute for Chinese and words per minute for English and Japanese. In Chinese, a non-demarcated script, reading rate was unaffected by the presence vs. absence of inter-word spaces.[41] In English, a demarcated script, reading rate decreased by approximately 50% without inter-word spaces.[20] In Japanese, a partially demarcated script, reading rate did not show significant improvement after the insertion of spaces between words.[8]

| | |
|---|---|
| Hindi | स्वच्छ हवा |
| Hebrew | אוויר יבש |
| English | Dry air |
| Arabic | الهواء الجاف |
| Thai | อากาศแห้ง |
| Korean | 건조한 공기 |
| Japanese | 乾燥した空気 |
| Chinese | 干燥的空气 |


Token frequency
Type frequency
Chinese
Japanese
Korean
Thai
Afrikaans
Arabic
Bulgarian
Catalan
Czech
Danish
English
Finnish
French
German
Greek
Hebrew
Hindi
Hungarian
Italian
Polish
Portuguese
Russian
Slovak
Slovenian
Spanish
Turkish
Ukrainian
Proportion
Word Length (characters/letters)


Ukrainian
Turkish
Spanish
Slovenian
Slovak
Russian
Portuguese
Polish
Italian
Hungarian
Hindi
Hebrew
Greek
German
French
Finnish
English
Danish
Czech
Catalan
Bulgarian
Arabic
Afrikaans
Thai
Korean
Japanese
Chinese
Written Languages
Information content of characters/letters
Information content of spaces
0.0
2.5
5.0
7.5
10.0
Information Content Contained by Spaces and Characters/Letters (bits)

With inter-word spaces
Without inter-word spaces
Reading Rate (words/min, characters/min)
200
100
0
239
239
283
159
261
119
88
78
108
120
Chinese
English: single line
English: passage
Japanese: Hiragana
Japanese: Kanji-Hiragana
Written Languages

**Table 1**

*Information of Different Written Languages*

| Written languages | Category | Example | Word length | Character/ Letter counts * | Information content of spaces | Information content of characters/letters |
|---|---|---|---|---|---|---|
| Chinese | None | 干燥的空气 | 1.40 (0.57) | 5623 | 1.10 | 9.84 |
| Japanese | Partial | 乾燥した空気 | 1.71 (0.78) | 2232 | 1.47 | 9.45 |
| Korean | Partial | 건조한 공기 | 2.48 (0.99) | 40 | 2.00 | 8.19 |
| Thai | Partial | อากาศแห้ง | 4.46 (2.77) | 102 | 3.21 | 5.60 |
| Afrikaans | Demarcated | Dry lug | 3.97 (2.24) | 42 | 2.84 | 4.20 |
| Arabic | Demarcated | الهواء الجاف | 3.89 (1.61) | 28 | 2.69 | 4.66 |
| Bulgarian | Demarcated | Suche powietrze | 4.25 (2.50) | 30 | 3.09 | 4.44 |
| Catalan | Demarcated | Aire sec | 4.01 (2.36) | 26 | 3.04 | 4.24 |
| Czech | Demarcated | Suchý vzduch | 4.26 (2.21) | 42 | 3.03 | 4.75 |
| Danish | Demarcated | Tør luft | 3.95 (2.06) | 29 | 2.77 | 4.27 |
| English | Demarcated | Dry air | 3.78 (2.04) | 26 | 2.90 | 4.32 |
| Finnish | Demarcated | Kuiva ilma | 5.52 (2.42) | 28 | 3.20 | 4.09 |
| French | Demarcated | L'air sec | 4.08 (2.31) | 26 | 2.93 | 4.32 |
| German | Demarcated | Trockene Luft | 4.58 (2.24) | 26 | 2.89 | 4.29 |
| Greek | Demarcated | Ξηρός αέρας | 4.57 (2.43) | 24 | 3.13 | 4.84 |
| Hebrew | Demarcated | אוויר יבש | 3.77 (1.42) | 32 | 2.47 | 4.43 |
| Hindi | Demarcated | स्वच्छ हवा | 3.42 (1.61) | 66 | 2.52 | 5.33 |
| Hungarian | Demarcated | Száraz levegő | 4.80 (2.58) | 46 | 3.30 | 4.59 |
| Italian | Demarcated | Aria secca | 4.47 (2.51) | 21 | 3.18 | 4.10 |
| Polish | Demarcated | Suche powietrze | 4.85 (2.62) | 32 | 3.31 | 4.59 |
| Portuguese | Demarcated | Ar seco | 4.33 (2.45) | 26 | 3.18 | 4.29 |
| Russian | Demarcated | Сухой воздух | 4.53 (2.62) | 33 | 3.27 | 4.61 |
| Slovak | Demarcated | Suchý vzduch | 4.26 (2.30) | 46 | 3.05 | 4.68 |
| Slovenian | Demarcated | Suh zrak | 4.25 (2.29) | 25 | 2.98 | 4.32 |
| Spanish | Demarcated | Aire seco | 4.34 (2.42) | 27 | 3.08 | 4.22 |
| Turkish | Demarcated | Kuru hava | 5.39 (2.52) | 29 | 3.27 | 4.43 |
| Ukrainian | Demarcated | Суший повітря | 4.32 (2.49) | 33 | 3.20 | 4.78 |

*Note.* The information content is measured in bits. Written languages can be categorized as demarcated, partially, or non-demarcated. All examples represent the phrase “dry air.” For word lengths, the means and standard deviations are shown.

* Information on Afrikaans, Bulgarian, Catalan, Czech, Polish, Slovak, Slovenian, and Turkish was retrieved from Wikipedia. Information on Chinese was retrieved from Lexicon

of Common Words in Contemporary Chinese. Letter count information was retrieved from Chang et al. for the other scripts.[53]