Myths about writing systems in speech & language technology

Kyle Gorman[*†] **and Richard Sproat**[†]
[*]CUNY Graduate Center
[†]Google LLC

Abstract

Natural language processing is largely focused on written text processing. However, many computational linguists tacitly endorse myths about the nature of writing. We highlight two of these myths—the conflation of language and writing, and the notion that Chinese, Japanese, and Korean writing is *ideographic*—and suggest how the community can dispel them.

1 Introduction

For a variety of historical and sociological reasons, *natural language processing* usually denotes the processing of *written* text, with work on spoken and signed language—as well as "multimodal" research—largely consigned to other venues. This largely unacknowledged focus on written language affects how computational linguists understand the nature of language itself. In this position paper, we argue that the field of natural language processing is beholden to certain misconceptions about the nature of writing and its relationship to other forms of language. We then make concrete suggestions as to how authors and editors can respond to these myths. We recognize that at least some of the issues we discuss may be obvious to the reader. If that is the case, we beg the reader's forgiveness. However, we think what we have to say here needs to be said.

2 Writing as a technology

Human language is notoriously hard to define, but we adopt a standard "cognitive" definition: the ability to learn and use systems of conventionalized externalized mental propositions. This ability is acquired, more or less effortlessly, by all typically developing humans barring gross sensory or motor impairments, and evolved sometime in the early prehistory of man. Writing, in contrast, is not a cognitive ability per se, but rather a technology which allows for the creation of durable visible records of language (Gelb, 1963, 11f.). Use of this technology can only be mastered by conscious, determined study[1] and has developed independently only a few times—in Mesopotamia and Egypt, China, and Central America—and in each case represents the dawn of human history in that region.

A writing system is, at its base, a linguistic analysis of the language it is used to write. (Indeed, the scribes of ancient Mesopotamia and Egypt are history's first linguists.) These analyses may seem quite naïve to the trained linguist, but the design of even the earliest writing systems hinge on sophisticated insights that presage the comparatively recent discoveries of the phoneme, mora, and morpheme.[2] For this reason, typologies of writing systems (e.g., Sproat, 2000; Rogers, 2005) are largely oriented around what types of linguistic units underlie the writing system's analysis.

3 Conflation and confusion

Language and writing may be ontologically incommensurate objects, but all known forms of writing are parasitic on spoken language. This is contrary to *standard language ideologies* (in the sense of Lippi-Green, 1997) which tend

[1]See, for example, Dehaene (2009) for an in-depth discussion of how parts of the brain which evolved for other purposes are co-opted in reading.

[2]Neural networks that process raw text—codepoint by codepoint or byte by byte—are sometimes said to work *from scratch* (e.g. Collobert et al., 2011). The implication seems to be that by doing away with explicit tokenization and related preprocessing steps, one has eliminated the need for linguistic analysis altogether. But since writing itself is a vernacular form of linguistic analysis, these could be said to work *from characters* or *from bytes* (e.g. Gillick et al., 2016; Li et al., 2019), but they certainly do not work from scratch.

Proceedings of the Workshop on Computation and Written Language (CAWL 2023), pages 1–5

to value written over spoken language, but it seems an unavoidable conclusion.

NLP researchers commonly conflate a language and the writing system(s) used to write it. Consider the following quotations, all taken from papers hosted on the ACL Anthology.[3]

"Right-to-left" As is well-known, Arabic, Hebrew, and Persian are written and read right-to-left.

> …*right to left* languages such as **Arabic** and **Hebrew**…
>
> Since **Persian** is a *right-to-left* language…

However, there is nothing about the **languages themselves** that is right-to-left. Furthermore, note that in Unicode, the right-to-left property of these scripts is purely an issue for text entry and rendering engines, since the codepoints are in the same logical order as text written left-to-right.

"Consonantal" Every language has consonants, so presumably the author below is referring to the consonantal alphabetic (or *abjad*) script used to write Arabic.

> One more idiosyncrasy of the **Arabic** language is that it is a *consonantal* language…

One does not ordinarily indicate short vowels in Arabic, except in certain pedagogical and religious texts. While the templatic word formation processes in Semitic languages might make them uniquely suited for this type of "defective" writing, many languages which lack this property—including Persian, Urdu, and until 1928, Turkish—are or were written using an Arabic-based consonantal script without great difficulty. This alone shows that there is nothing particularly "consonantal" about the Arabic language.

"Syllabic" Much like the presence of consonants, division of spoken language into syllables seems to be a linguistic universal, so it is not clear why the authors quoted below have chosen to highlight this property.

> …**French** is a *syllabic* language…
>
> …**Linear B**, a *syllabic* language…
>
> **Mandarin** is a tonal and *syllabic* language…
>
> **Punjabi** is a *syllabic* language…

Indeed, Punjabi is notable for having two major—and rather different—writing systems, the Gurmukhi alphasyllabary (or *abugida*) and the Shahmukhi consonantal alphabet. Neither of these systems use the phonological syllable as an orthographic unit, though alphasyllabaries have been characterized as using so-called *orthographic syllables*.

Chinese Finally, one can find a number of conflicting statements about the nature of Chinese in the ACL Anthology.

> **Chinese** is a *morphemic* language.
>
> **Chinese** is a *logographic* language…
>
> …**Chinese** is *ideographic*
>
> It's well known that **Chinese** is an *ideographic* language…
>
> …**Chinese**, **Japanese** and other *ideographic* languages.

We now turn to the question of what kind of writing system Chinese really is.

4 Ideography and CJK

We acknowledge that there exist many symbol systems which are purely *ideographic* (or *semasiographic*), without any direct reference to spoken language (Sproat, 2023). However, DeFrancis (1989, ch. 2) shows that these fail to satisfy any reasonable definition of writing. There have also been heroic attempts to develop purely ideographic writing systems, most notably Blissymbolics (Bliss, 1965). While carefully designed, such systems struggle with encoding categories like:

- colors; e.g., *chartreuse*, *royal blue*
- proper names; e.g., *Kyle*, *Richard*, *Park Slope*, *Shibuya*
- non-imageable predicates; e.g., *imagine*, *consternation*

[3] We deliberately omit citations for these quotations. We do not wish to draw undue negative attention to particular authors, but simply to illustrate how widespread this confusion is within our community.

- subtle connotative differences; e.g., *salt* vs. *sodium chloride* vs. *NaCl*.

Dependency on spoken language appears to be inherent to the design of writing. Despite this, it is very common to find statements in the literature that suggest that some writing systems, especially those used for Chinese, Japanese, or Korean, are *ideographic*. We demonstrate this with a survey of the literature in subsection 4.4, but first we present a brief synopsis of how Chinese, Japanese and Korean writing actually work.

4.1 Chinese writing

As argued by DeFrancis (1989, ch. 3), Chinese writing is best described as *morphosyllabic*. With only rare exceptions, each Chinese character represents a single phonological syllable, and in most cases corresponds to a single morpheme as well. Some characters, such as 人 *rén* 'person' are nondecomposable in that they represent the respective morpheme, and cannot be broken down into parts that have any meaning on their own. However *most* characters that have ever been invented—roughly 90%, by some estimates—are *semantic-phonetic* compounds that can be decomposed into a portion that represents something about the meaning and another portion that represents something about the pronunciation. For example, 鯉 *lǐ* 'carp', can be broken down into 魚, meaning 'fish' and 里, which here is being used for its pronunciation *lǐ*.[4] In most cases the pronunciation hint provided by the phonetic component is not nearly as good as in the case of 鯉, but the crucial point is that despite the common myth that Chinese writing is *ideographic*, in fact it depends heavily on phonology.

4.2 Japanese writing

The adaptation of Chinese writing to Japanese is more complex. In Japanese, *kanji* are used both to represent words or morphemes of Chinese origin, as well as native words. In the former case, the same semantic-phonetic principles carry over in that the phonetic component serves the purpose of hinting at the (Sino-Japanese) reading of the morpheme. In the latter case, the Chinese phonetic component is generally useless. For example the reading of 鯉 'carp' in Japanese is *koi*. In such cases, the kanji comes unanalyzable, like 人 'person' in Chinese, in that there is no phonetic cue to the reading (though the semantic component still has some function): 鯉 is just used as a whole to represent the morpheme *koi*. But notice that the unit represented is still a linguistic unit—a morpheme—not an idea (cf. Joyce, 2011). Apart from kanji, a large portion of Japanese writing is covered by *hiragana* and *katakana*, two (moraic) syllabaries that were historically derived from using Chinese characters purely for their pronunciation, and are now reasonably transparent, phonemic systems. That said, there are cases where the Japanese use of kanji is nearly semasiographic. One case is where different kanji are used to spell different senses of the same etymon. For example 泊まる *tomaru* 'stay, stop at a lodging' is probably the same word as 止まる *tomaru* 'stop, come to a halt', but the two spellings reflect different senses. Another case involves *jukujikun*, native Japanese words that are written with multiple kanji purely for their meaning. For example *sanma* 'Pacific saury', has a kanji spelling 秋刀魚 whose individual kanji convey the meaning 'autumn sword fish'. Since saury are long silver fish usually caught in the autumn, this spelling certainly evokes the meaning, but none of the kanji individually correspond to any linguistic unit. However, this reflects a small portion of the writing system, and the vast majority of Japanese writing can still be characterized as phonological or morphological.

4.3 Korean writing

Chinese characters were used widely in the history of Korean. Adaptations of Chinese writing to Korean were very similar to what one finds in Modern Japanese; see Handel (2019) for an in-depth discussion. In contrast, Modern Korean makes very sparing use of Chinese characters, and then only for morphemes of Chinese origin; e.g., 男 *nam* 'male' and 女 *yeo* 'female' on bathroom signs, or 小 *so* 'small', 中 *jung* 'medium', and 大 *dae* 'large' for serving sizes in restaurants. Nearly all Korean text

[4] There are also quite a few characters that are decomposable into two or more bits that represent the meaning, but unlike semantic-phonetic compounds, these have not been a major source of new characters in the last few millennia.

nowadays is written in *hangul*, the alphabetic writing system developed in the 15th century under King Sejong. Korean writing is thus essentially phonological.

4.4 Methods

Since the ideography myth is so pervasive in the speech and language processing community, it seems useful to try to understand what authors wish to convey when they incorrectly describe a language or writing system as *ideographic*. Therefore, we conducted an exhaustive survey of the ACL Anthology[5] for the words *ideograph*, *idiograph* [sic], and *ideographic*. There is little speech processing work published in the Anthology, and there is no single central repository for research on this topic. To survey speech research, the search terms *"ideographic" "speech recognition"* and *"ideographic" "speech synthesis"* were entered into Google Scholar.[6] As anticipated, this procedure retrieved examples from the proceedings of conferences like ICASSP, INTERSPEECH, and ASRU, and journals like *Computer Speech & Language*. ACL Anthology papers were excluded from this latter sample.

50 papers, all published 2003–2022, were selected randomly from each of the two samples. We then examined the surrounding context in which ideography is mentioned, and manually coded the following:

- which languages and/or writing systems this term refers to,
- whether or not language and writing is conflated, and
- the authors' apparent reason for mentioning ideography.

One author [KG] coded the Anthology sample, and another [RS] coded the Scholar sample.

4.5 Results

Three general trends emerge. First, as shown in Table 1, Chinese and Japanese are by far the most common languages to be described as ideographic. Three sources described Korean as ideographic. As we noted above (subsection 4.3), modern Korean writing makes limited use of Chinese characters, but it is not clear why this is qualitatively different than, for example, English's ideographic use of the dollar and pound signs in currency expressions like *$4.20*. Akkadian cuneiform and Egyptian hieroglyphs are both mentioned; these would probably be regarded as mixed writing systems, (cf. Hermalin this volume and Sproat and Gutkin 2021 for recent attempts to quantitatively characterize such scripts).

	Anthology	Scholar
Chinese	31	37
Japanese	20	21
(others)	8	1

Table 1: The counts of languages and/or writing systems described as "ideographic" (etc.) in a sample of 100 speech & language processing papers, published 2003–2022 in either the ACL Anthology or in speech research venues via Google Scholar.

More bafflingly, the undeciphered Proto-Elamite script, known to us from Early Bronze Age inscriptions in Iran, is similarly described, as is Dutch, and the entire Indo-Aryan language family. Secondly, 23 of the 100 papers are quite explicit in incorrectly conflating writing and language (i.e., describing Chinese as an *ideographic language*). Third, 69 of the 100 papers which mention ideography appear to do so as a means to describe—or simply introduce—the Han characters used in Chinese, Japanese, and Korean. However, we note some correctly describe symbols such as $, &, and Arabic numbers as *ideographic*.

5 Conclusions

We have shown that researchers in speech and language processing frequently conflate writing and language, a mistake that is often accompanied by misunderstandings about the nature of writing itself or misinformation about the nature of specific writing systems.

We recognize that many researchers may lack the necessary background in writing systems, and this is unsurprising given that the history and structure of writing is not widely taught, at least at North American universities. For researchers who wish to learn more about writing systems, we recommend two texts: Rogers (2005) provides an accessible introduction to the typology of writing sys-

[5] https://aclanthology.org/
[6] https://scholar.google.com

tems, and Gnanadesikan (2009) gives an easy-to-read introduction to the history of writing.

Editors and reviewers should pay more attention the use of inappropriate terminology used to describe writing systems as a simple matter of scientific communication. Describing the Arabic and Chinese languages as *right-to-left* or *ideographic* wrongly conflates of writing and language; these are the sort of mistakes that simply should not be made by specialists in our field.

We conclude with one concrete suggestion: we recommend the Unicode Consortium remove incorrect uses of the term *ideograph* in the standard (Unicode Consortium, 2021). As the they admit in their FAQ on CJK languages, this term does not accurately reflect the nature of these characters, but they claim the "term is now so pervasive in the standard that it cannot be abandoned or replaced."[7] However, we submit that the standard's description of Han characters as *CJK Unified Ideographs* is perhaps the primary channel by which the myth has been propagated amongst technologists, and a correction and *mea culpa* would do much to publicize the issue and dispel this myth. Handel (2019), for instance, proposes the term *sinographs*, and there are other sensible alternatives.

Acknowledgments

Thanks to Brian Roark, Daniel Yakubov, and audiences at Grapholinguistics in the 21st Century for comments on this work.

References

Charles K. Bliss. 1965. *Semantography (Blissymbolics)*, 2nd enlarged edition. Semantography (Blissymbolics) Publications.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

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

Stanislas Dehaene. 2009. *Reading in the Brain: The Science and Evolution of a Human Invention*. Viking.

I. J. Gelb. 1963. *A Study of Writing*, 2nd edition. University of Chicago Press.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306.

Amalia E. Gnanadesikan. 2009. *The Writing Revolution: Cuneiform to the Internet*. Wiley-Blackwell.

Zev Handel. 2019. *Sinography: The Borrowing and Adaptation of the Chinese Script*. Brill.

Noah Hermalin. 2023. A mutual information-based approach to quantifying logography in Japanese and Sumerian. In *Proceedings of the ACL Workshop on Computation and Written Language*.

Terry Joyce. 2011. The significance of the morphographic principle for the classification of writing systems. *Written Language and Literacy*, 14(1):58–81.

Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, and William Chan. 2019. Bytes are all you need: end-to-end multilingual speech recognition and synthesis with bytes. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5621–5625.

Rosina Lippi-Green. 1997. *English with an Accent: Language, Ideology, and Discrimination in the United States*. Routledge.

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

Richard Sproat. 2000. *A Computational Theory of Writing Systems*. Cambridge University Press.

Richard Sproat. 2023. *Symbols: An Evolutionary History from the Stone Age to the Future*. SpringerNature.

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

Unicode Consortium. 2021. *The Unicode® Standard: Version 14.0: Core Specification*. Unicode Consortium.

[7] https://unicode.org/faq/han_cjk.html