The persistent conflation of writing and language

Kyle Gorman[*†] & Richard Sproat[†]

[*]Graduate Center, City University of New York

[†]Google

Grapholinguistics in the 21st Century

Caveat

- Our remarks today are directed at the speech and language processing community. **What we have to say needs to be said.**
- But some of the points we will make may be obvious to our fellow grapholinguists.

Definitions

- By *language* we mean the ability to externalize complex mental propositions, evolved by *H. sapiens* sometime during the Middle Stone Age.
  - This ability is acquired more or less effortlessly by all typically developing humans, barring gross sensory or motor impairments.
- By *writing* we mean the technology which allows us to create discrete, durable physical records of spoken language (Gelb 1952).
  - This technology can only be mastered by conscious, determined study.
  - It developed only a few times independently:
    - c. 3000 BCE in Mesopotamia and Egypt
    - c. 1500 BCE in China
    - c. 300 BCE in Central America

NLP as written language processing

For a variety of historical and sociological reasons, nearly all natural language processing (NLP) research involves processing of text—that is, written documents–with work on spoken and signed language (as well as much "multimodal" work) largely relegated to other venues. So, NLP is almost exclusively *written language processing*.

This–largely unacknowledged–focus on written language leads substantial confusion among NLP practitioners, very few of whom will have studied the world's writing systems in any detail.

NLP researchers should clearly and explicitly differentiate between language and writing.

Proscriptions

Thus, NLP researchers should not conflate

- language and writing;
- a language and a script; nor
- the properties of a script with properties of the language it is used for.

One possible origin for this conflation are *standard language ideologies* (in the sense of Lippi-Green 1997) which view written language as superior to–if not also logically prior to–spoken language.

Grammatology

Certain scripts may have affordances for writing particular types of languages:

- Consonantal alphabets may be well suited to represent templatic word formation in Semitic.
- The earliest morphographic scripts evolved for languages in which most stems were monosyllabic (cf. Sproat 2017).

But these linguistic properties are not necessary conditions, because, e.g., consonantal alphabets and morphographic scripts have been used for dozens of languages without these properties.

What not to say

All of the following examples are taken from the ACL Anthology, with citations hidden to protect the accused…

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

- These scripts are read right-to-left, but when working with Unicode (or UTF-8) text, this is a merely a property of the rendering system, since their codepoints (or bytes) are in the same logical order as any other text.
- There is nothing about the language itself that is "right-to-left".

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

- Every language has consonants, so presumably this is referring to consonantal alphabetic (or *abjad*) script used to write Arabic, which does not write short vowels (except in certain religious and pedagogical texts).
- While templatic word formation in Semitic languages may make them uniquely suited for this type of defective writing, dozens of languages which lack these properties are written using this script.

“**Punjabi** is a *syllabic* language…”

- Every language has syllables, so presumably this is referring to the Gurmukhi alphasyllabary (or *abugida*) used to write Punjabi in India.
    - Alphasyllabaries are *not* syllabaries, and their *orthographic syllables* do not correspond to phonological syllables.
- Of course, in Pakistan it's largely written using the Shahmukhi consonantal alphabet, which isn't syllabic in any relevant sense.

“CJK” writing

For whatever reason, this conflation is particularly common when talking about Chinese, Japanese, and Korean. E.g.:

“**Chinese** is a *morphemic* language.”
“**Mandarin** is a tonal and *syllabic* 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."

The limits of ideography

- All writing systems depend in part on a—possibly naïvely, possibly quite sophisticated—linguistic analysis that is in part phonological and/or morphological (e.g., DeFrancis 1989, Sproat 2010).
- It is likely *impossible* to construct a purely *ideographic* symbol system that would satisfy any meaningful definition of writing.

Challenges for ideography

Some problems include the encoding of:

- Proper names; e.g., *Kyle, Richard*, *Park Slope, Shibuya*
- Colors; e.g., *chartreuse*, *royal blue*, *cerulian*
- Non-imageable predicates; e.g., *imagine, freedom, consternation*
- Subtle connotative differences; e.g., *salt* vs. *sodium chloride* vs. *NaCl*

The limits of iconography/ideography (after Sproat 2010)

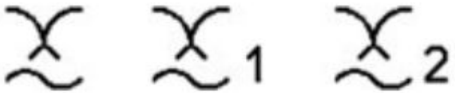


Figure 2.8 Use of indices in Blissymbolics: 'waterfowl', 'duck' (waterfowl + 1), 'goose' (waterfowl + 2)

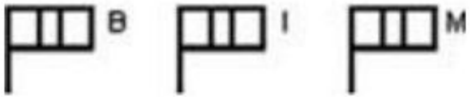
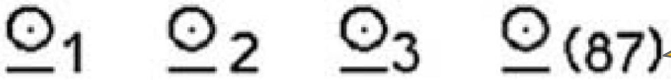


Figure 2.10 Use of letters in Blissymbolics: 'Belgium', 'Italy', 'Mexico'



This one is especially easy to remember 🙄

Figure 2.17 Some colors in Blissymbolics: 'red', 'orange', 'yellow', 'Persian blue'

The survey (part 1)

- We wanted to understand what authors mean when they say "ideograph(ic)".
- We conducted an exhaustive survey of these terms in the ACL Anthology (since 2003). We searched for *ideograph*, *idiograph* [sic], and *ideographic*, coding:
  - which languages or writing systems this term refers to,
  - whether or not language and writing is conflated,
  - and the authors' apparent reason for mentioning this notion.

Link to the survey results

ACL Anthology survey results (50 cases)

- Most commonly described as "ideographic":
  - Chinese (31)
  - Japanese (20)
  - Korean (2)
  - Others: Akkadian, Blissymbols, Dutch (!), Egyptian, Indo-Aryan (!), Proto-Elamite
- Conflating writing and language (13)
- Most common reasons to mention ideography:
  - As a description of Han characters (29)
  - To motivate word or subword segmentation methods (5 + 5)
  - Unclear (7)

The survey (part 2)

- Whereas virtually all NLP research appears in the ACL Anthology, there is no similar repository for speech processing research, so we instead searched Google Scholar for the terms:
    - "ideographic" "speech recognition"
    - "ideographic" "speech synthesis"
- This found similar examples from the proceedings of conferences like ICASSP, INTERSPEECH, ASRU, etc., from 2003 onwards.

ASR and TTS survey results (50 cases)

- Most commonly described as "ideographic":
  - Chinese (37)
  - Japanese (21)
  - Korean (1)
- Conflating writing and language (10)
- Most common reasons to mention ideography:
  - As a description of Han characters (40)

Summary

- "Ideograph" and similar terms are often used simply to describe or introduce the scripts which use Han characters.
- About 20% of the time there is a clear conflation of language and writing.
- In a few cases symbols such as $, &, 1, 2, 3, etc. are *correctly* described as "ideographic".

Suggestions (1/)

- The time has come for the Unicode Consortium to repent and make amends for their abuse of the term *ideograph*. From their CJK FAQ:

  > Q: Why does Unicode use the term "ideograph" when it is linguistically incorrect? The characters used to write Chinese are…generally referred to by names such as "ideograph" or "pictogram," even though these don't accurately reflect what the characters are or how they are used. [...] Unicode originally adopted the word "ideograph" as representing common English usage. **The term is now so pervasive in the standard that it cannot be abandoned.**

- But describing Han characters as “CJK Unified Ideographs” perpetuates a myth.
- One possibility is to replace it with *sinograph* (e.g., Handel 2019).

Suggestions (2/)

- Editors, area chairs, and reviewers need to pay more attention and catch inappropriate terminology when it arises.
  - The authors have caught many such issues in paper reviews in the past.
  - But researchers should be made aware of these issues.
- Is it time for a special interest group on writing systems?
  - There is some real interest in the computational analysis of writing systems.
  - If this was associated with the ACL, it would at least get the attention of the NLP community.

Suggestions (3/)

- There is a recent trend for "from scratch" neural network systems which work at the level of individual Unicode codepoints or even UTF-8 encoded bytestrings (see, e.g., Gillick et al. 2016, Li et al. 2019).
    - There is an implication that "from scratch" approaches somehow eliminate the need for linguistic insight altogether.
    - But writing systems are a type of linguistic analysis, and while their analyses may be quite naïve, they encode sophisticated phonemic and/or morphemic insights in symbolic form.
- Such models should not be described as working "from scratch".

Questions?

Selected bibliography

DeFrancis, J. 1989. *Visible Speech: the Diverse Oneness of Writing Systems*. University of Hawaii Press.

Gelb, I. J. 1952. *A Study of Writing*. University of Chicago Press.

Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. 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.

Handel, Z. 2019. *Sinography: the Borrowing and Adaptation of the Chinese Script.* Brill.

Li, B., Zhang, Y., Sainath, T., Wu, Y., and Chan, W. 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.

Selected bibliography

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

Sproat, R. 2010. *Language, Technology, and Society*. Oxford University Press.

Sproat, R. 2017. A computational model of the discovery of writing. *Written Language & Literacy* 20: 194-226.

Unger, J. M. 2004. *Ideogram: Chinese Characters and the Myth of Disembodied Meaning*. University of Hawaii Press.

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