Computational Methods in the Analysis of Graphical Symbol Systems

Richard Sproat
Google, Japan
rws@google.com

Grapholinguistics in the 21st Century
Télécom Paris, Palaiseau
June 8, 2022

Background

- Humans have been using graphical symbols for thousands of years.
- Most of these have not been tied to language.
- But about 5,000 years ago a special form of graphical symbol system evolved that was *intimately* connected to language: writing
- **How can quantitative/computational models help us understand various aspects of symbol systems, writing systems and their evolution?**

Some problems

I will explore some *computational* approaches to the following questions:

- What does it mean to say a writing system is *logographic*?
- You have an “undeciphered” symbol system.
  - Was it writing or some sort of non-linguistic system?
    - Part of the problem is being clear on what you mean by terms like “writing”.
    - And, assuming we are clear on *that* point, does “structure” imply it’s writing?
- How did writing evolve from non-writing? In this case a lack of clear archaeological evidence could be supplemented with computational approaches. [Synopsis only]

Computational methods force you to be specific about your assumptions.

Logographic Writing

(joint work with Alexander Gutkin*)

*Sproat & Gutkin (2021)

What does the term “logography” refer to?

- There have been a lot of definitions of “logography” (“morphography”) – mostly imprecise
- A few examples:
  - 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.”
  - Sampson (1985, p. 33): “logographic systems are those based on meaningful units”
  - 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.”
  - 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**.”
- Handel’s definition is the closest to the one we adopt, though he actually means something far more specific, namely the kind of logographic elements familiar from Chinese writing.

Classifying writing systems

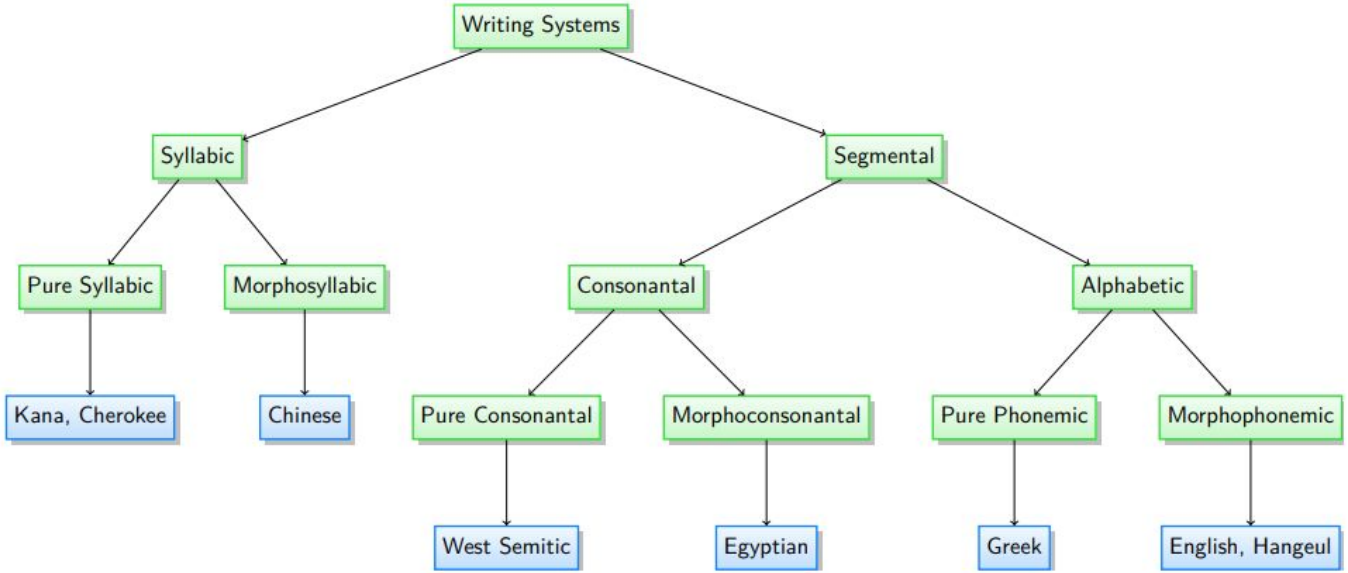


Figure 1
DeFrancis' (1989) taxonomy of writing systems; simplified from his Figure 10, page 58, to focus on what he considers to be true writing systems.

Classifying writing systems

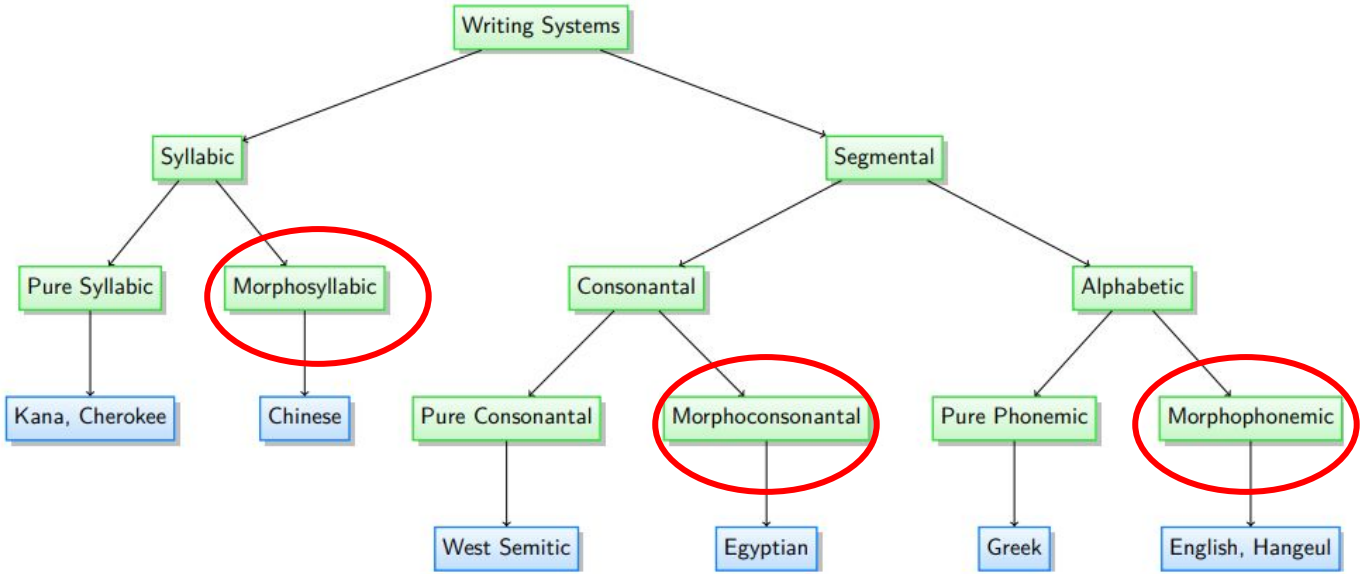


Figure 1
DeFrancis' (1989) taxonomy of writing systems; simplified from his Figure 10, page 58, to focus on what he considers to be true writing systems.

Classifying writing systems

More or less categorical – though see Osterkamp & Schreiber (2021)

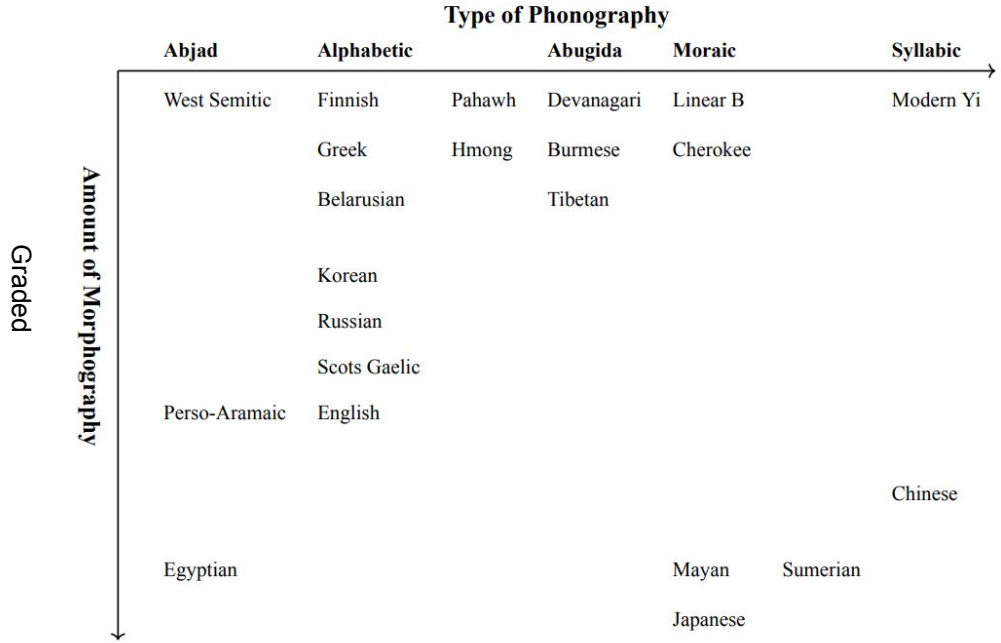


Figure 2
Rogers' (2005) planar taxonomy (his Figure 14.5, page 275), developed based on an earlier proposal in Sproat (2000) (= Sproat's Figure 4.5, page 142).

Classifying writing systems

More or less categorical – though see Osterkamp & Schreiber (2021)

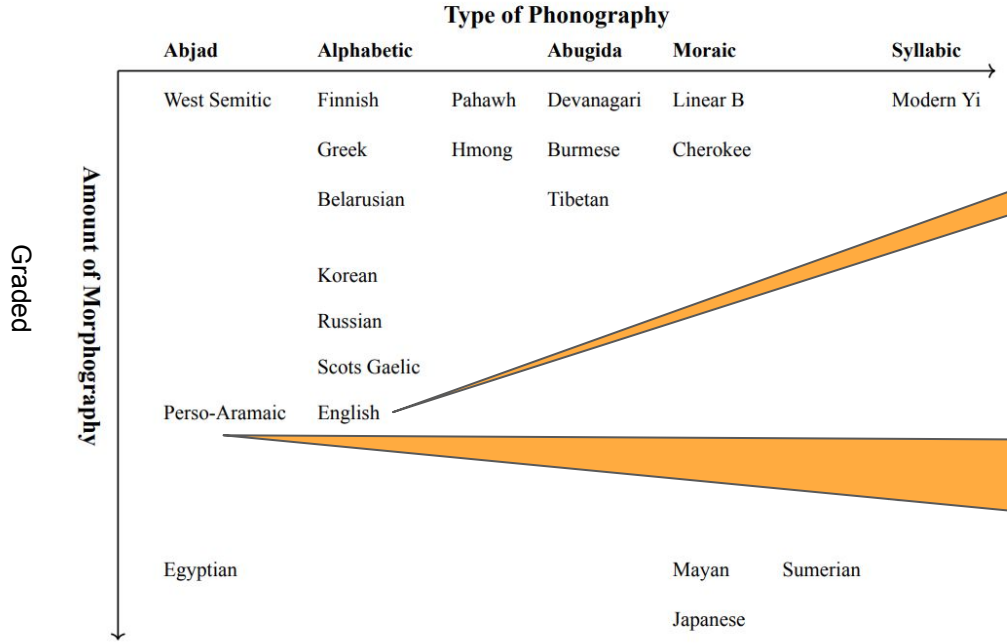


Figure 2
Rogers' (2005) planar taxonomy (his Figure 14.5, page 275), developed based on an earlier proposal in Sproat (2000) (= Sproat's Figure 4.5, page 142).

Cf. Sampson (1985)

heterograms of Persian *morphemes* spelled as their Aramaic translations

logography isn't only morpheme-sized "graphemes"

Logography (morphography) is a matter of degree

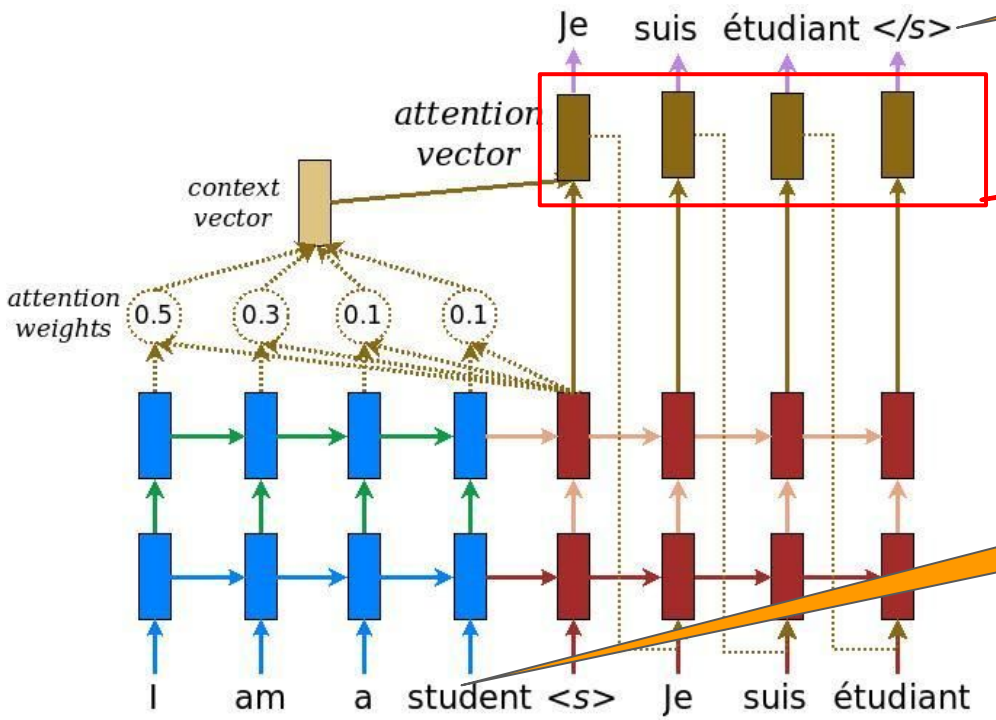
- If it is a matter of degree then one ought to be able to measure it
  - Sproat (2000) didn't provide such a measure. Nor did Rogers (2005)
  - Penn & Choma (2006) proposed using *correlation coefficients*.
    - Doesn't work (Sproat & Gutkin, 2021)
- **A measure that works well is the amount of *attention* paid to the *context* by a neural sequence-to-sequence model trained to *spell* words in context.**

The task: spell a word in context

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

Attention-based RNN model



In our case: output is "graphemes"

Combine attention vectors into attention *matrix* visualization

In our case: input is phonemes

See the paper for details of the network and how it was trained.

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

Attention in a highly phonemic writing system

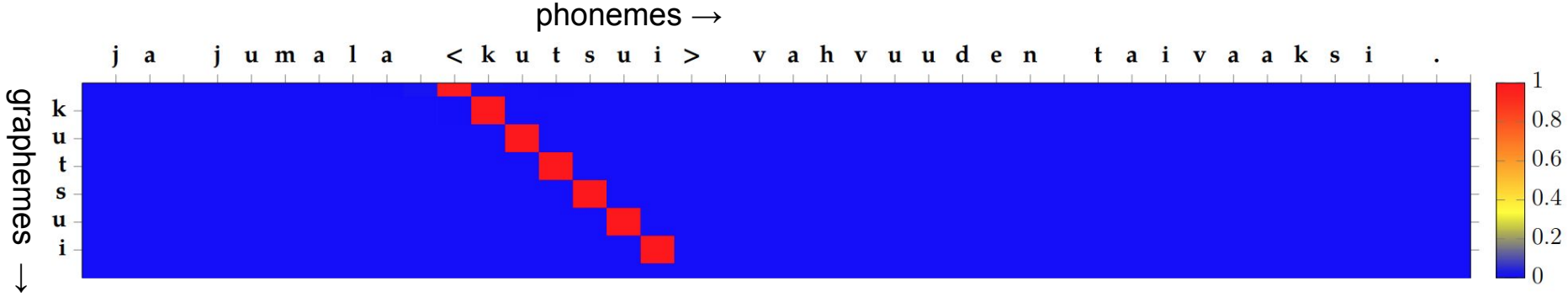


Figure 7
Attention matrix involved in spelling the Finnish word *kutsui* 'called'. The input (phonetic) sequence for the sentence is shown across the top of the plot, and the spelling of the target word is shown on the vertical axis. Note that in the plot itself the `<targ> ... </targ>` tags are reduced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

Attention in a highly logographic writing system

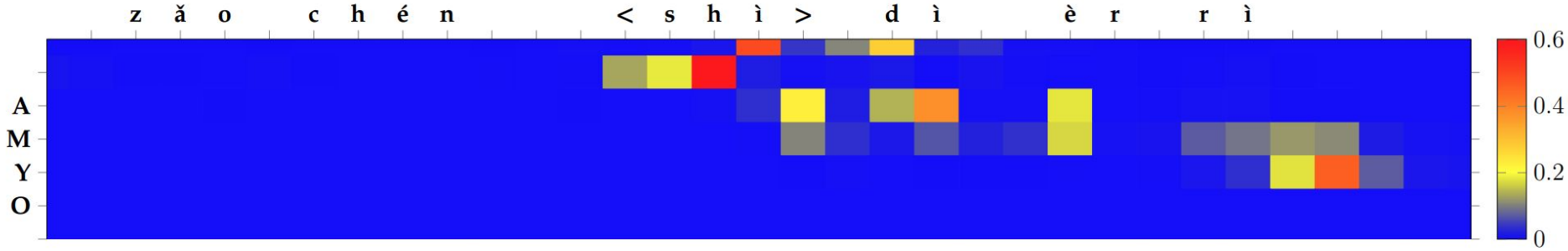


Figure 8

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.

Attention-based measure

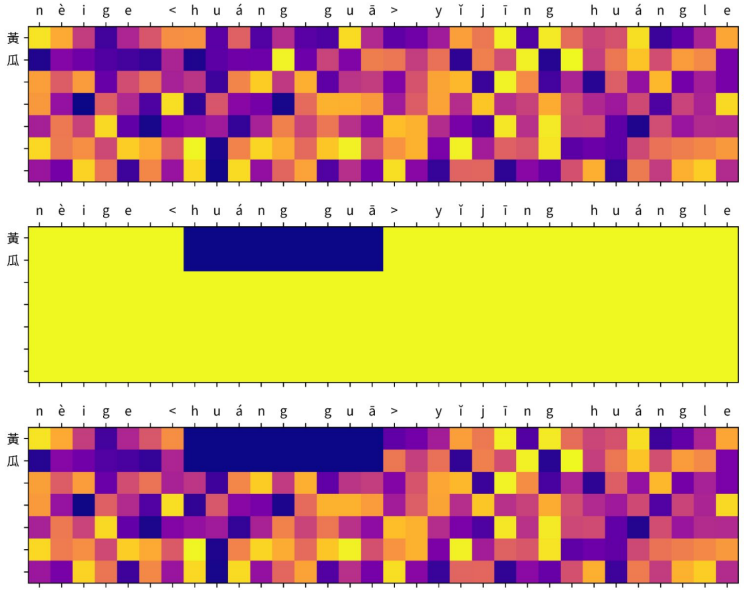


Figure 9
Illustration of the attention-based spread measure. Top: A random attention matrix. Middle: The zero mask for the target word. Bottom: The Hadamard product of the mask with the attention matrix.

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

A is the attention matrix
M is the mask

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

N is the size of the corpus
V is the size of the vocabulary
|*v*| is the number of instances of type *v*

Logography (morphography) is a matter of degree

- We also considered various other measures including:
  - Lexical measure – based on the spelling variants for a given phonological form in a dictionary (see also Marjou, 2021).
  - Entropic measure – compare bigram conditional entropy of the written form vs. phonology.
  - Both of these have *type* and *token* interpretations.

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \log p(y|x)$$

- Where *X* and *Y* are random variables and p(*y*|*x*) is the probability of *y* following *x.*
- Main point:
  - *higher* entropy → *less* predictable, less information in context.
  - *lower* entropy → *more* predictable, more information in context.

Data: Bible corpus

Old Testament only.
Written side is undiacritized.
Modern/Biblical prons
derived from diacritization.

Jamo
(individual
hangul letters)

Variants:
tokenized +/-
cangjie +/-

Variants:
cangjie +/-

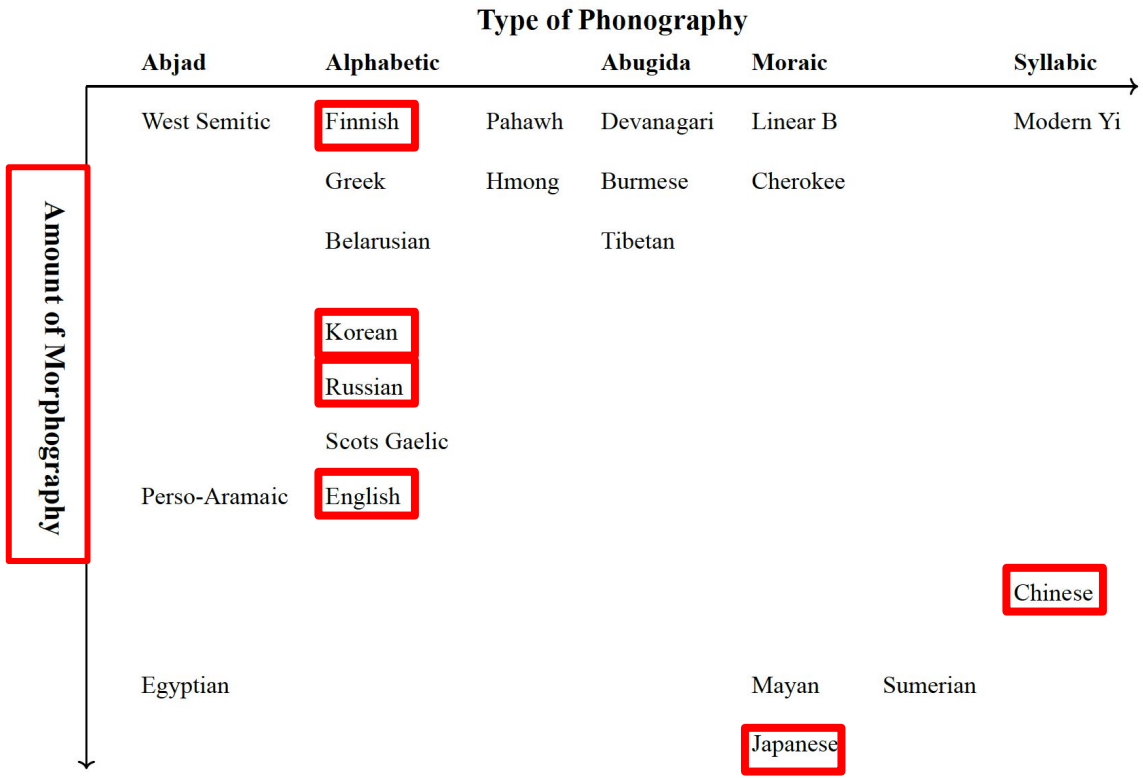
Table 3
Summary of the resources used for each of the language

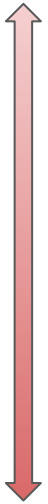
Language	Phonetic Transcription	Addi… packages/sources used
English	ARPAbet	…s://pypi.org/project/pronouncing/
French	Idiosyncratic system	…ttp://www.lexique.org/databases/Lexique3…
Russian	Idiosyncratic system	https://github.com/kylebgorman/wikipron
Finnish	Finnish letters	
Swedish	SAMPA-derived	http://www.nb.no
Hebrew (Biblical) Hebrew (Modern)	Idiosyncratic system Idiosyncratic system	https://www.mechon-mamre.org
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

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 つ/tsu た/ta 。/.

Rogers' planar taxonomy again





Finnish

Korean

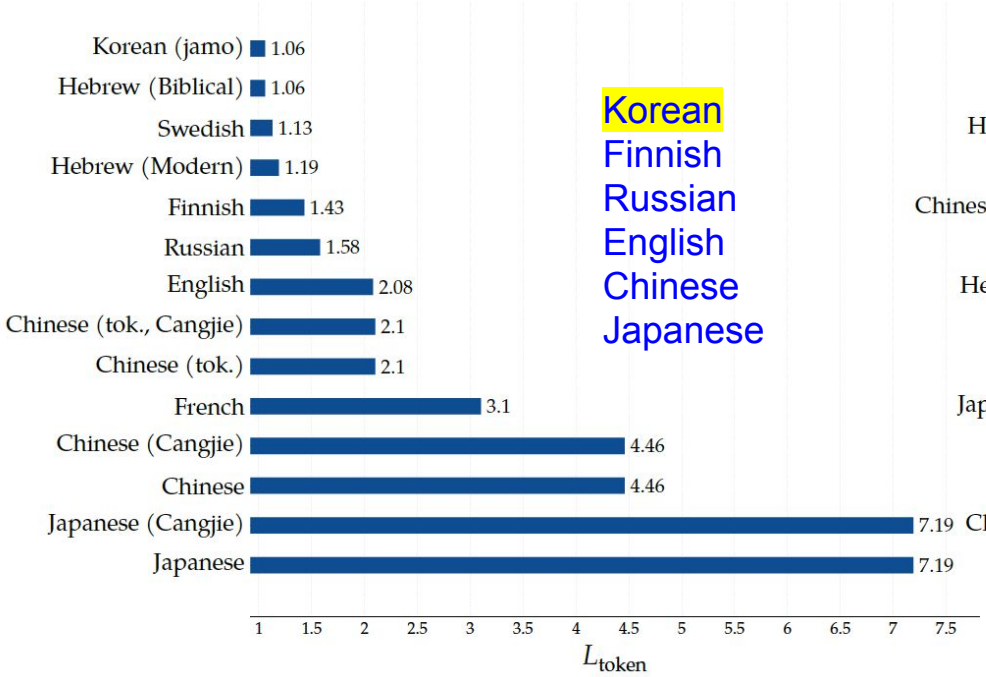
Russian

English

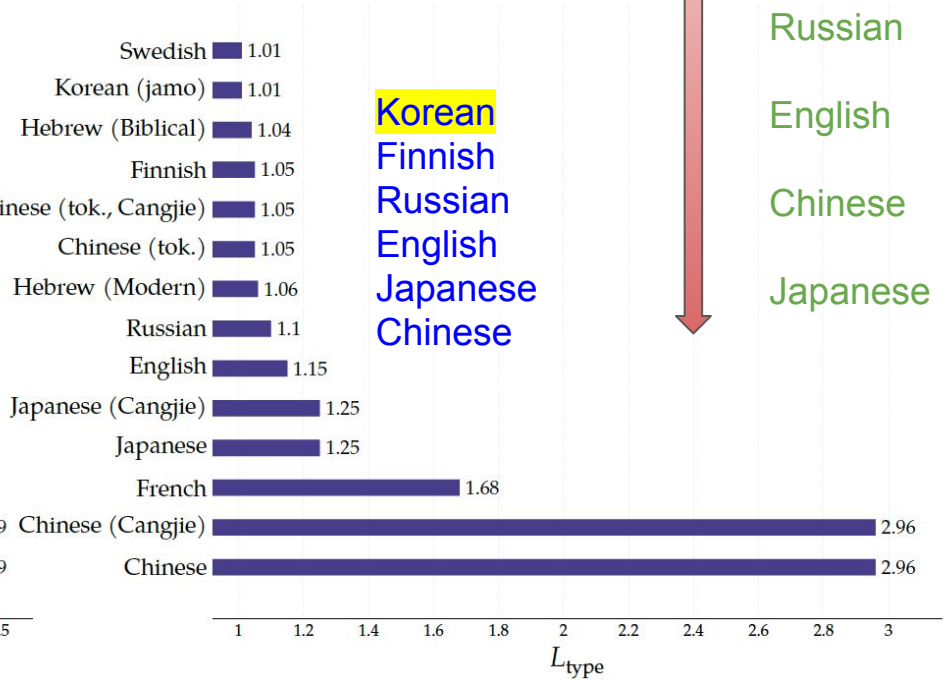
Chinese

Japanese

Results: Lexical measures



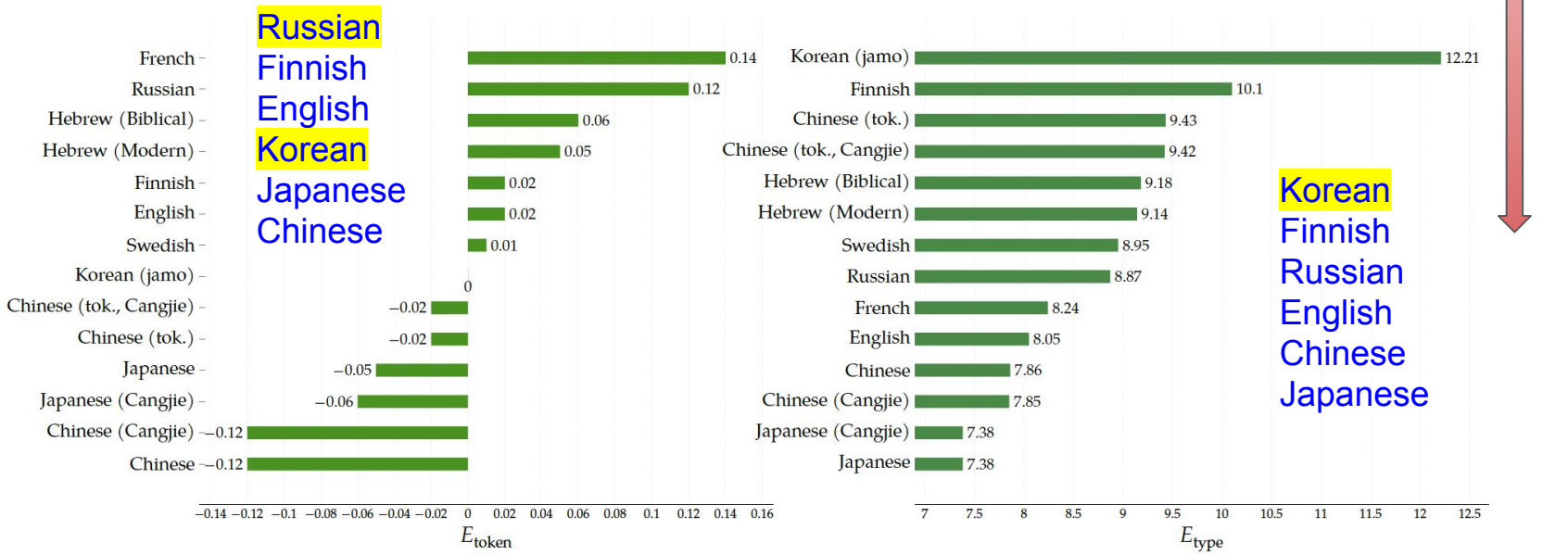
Korean
Finnish
Russian
English
Chinese
Japanese



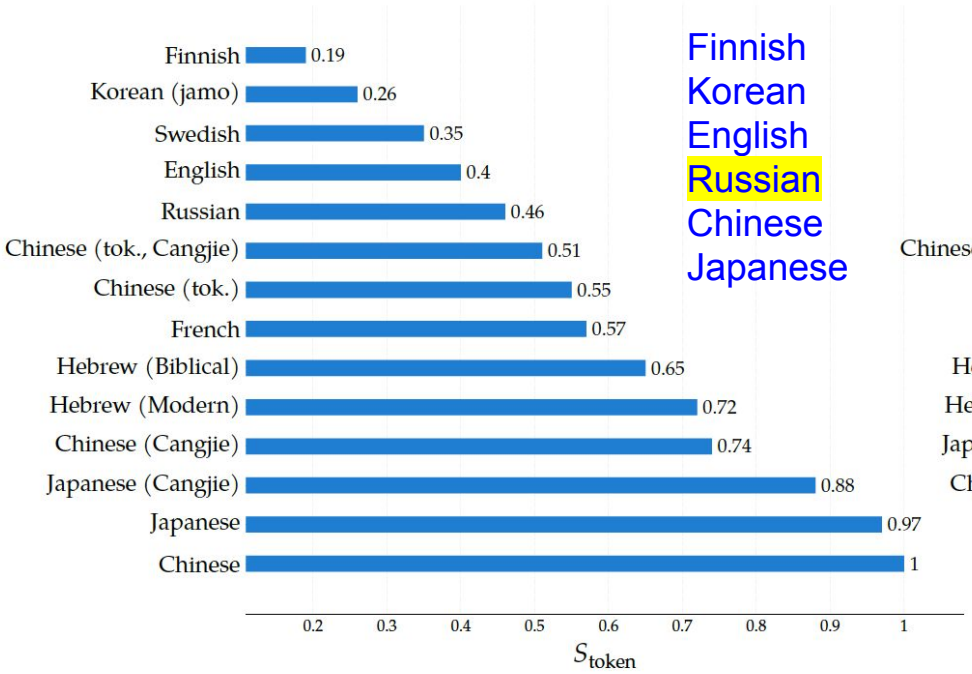
Korean
Finnish
Russian
English
Chinese
Japanese

Finnish
Korean
Russian
English
Chinese
Japanese

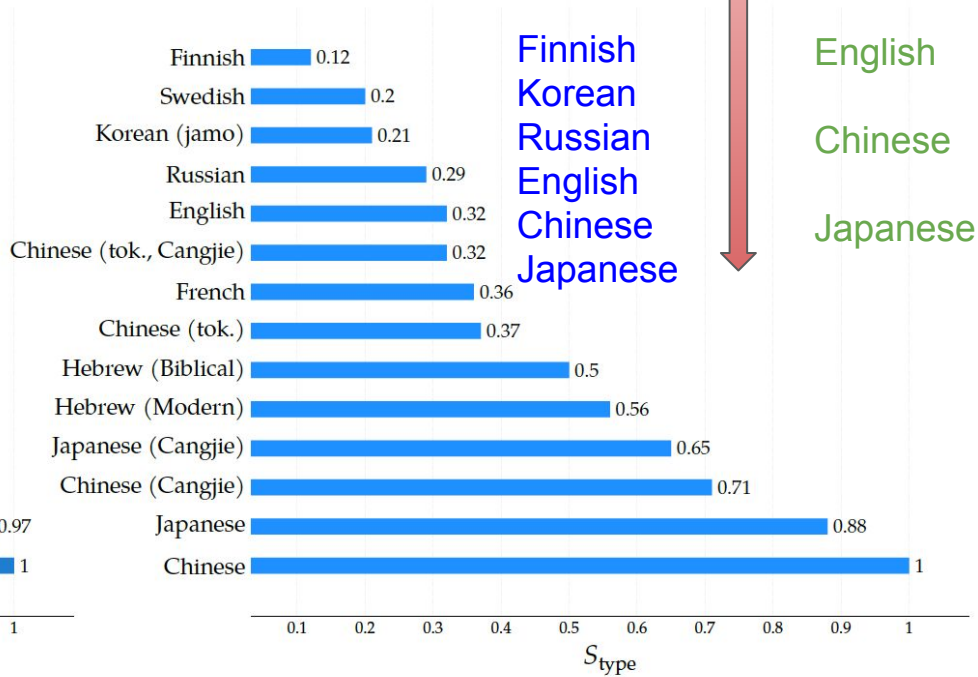
Results: Entropic measures



Results: Attention-based measures



Finnish
Korean
English
Russian
Chinese
Japanese



Finnish
Korean
Russian
English
Chinese
Japanese

Synopsis

- Compared to measures based on simple lexical counts of spelling alternatives, or entropy, the attention-based measure seemed to accord better with intuition.
- More sophisticated neural models – transformers, temporal convolutional models – were about the same … but much harder to interpret.
- We also argue in the paper that the results correlate with comparative studies of the acquisition of spelling.
- Conclusion:
  - One *can* measure the degree of logography in a way that seems to accord with intuition.
  - The model makes intuitive sense:
    - The more logographic a system is, the more context is needed to predict spelling
  - Logography isn't binary: some writing systems are more logographic than others.
  - Amount of logography also depends on *what* you define the target to be (morpheme, word…)

Writing versus Non-Writing
(How computational methods can be misused)

Two recurring problems

- Fuzzy or (intentionally?) unclear definitions of the notion “writing”.
- The fallacy that *structure* (AKA *syntax*) ⇒ *linguistic structure.*
- Computational methods force you to be precise on these points and expose you when you aren’t.

What is writing and how do you tell if you've got it?

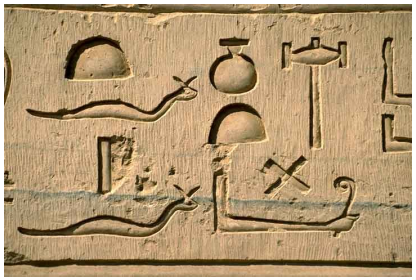
- How do you tell if an *uninterpretable* symbol system is writing? *Can* you tell?
  - If there are only single symbols (no texts); or symbols repeat too much; or symbols never repeat … then you'd probably guess it isn't

Pennsylvania “barn stars” (AKA “hex signs”)

Farrell collection, Berks County Historical Society

What is writing and how do you tell if you've got it?

- How do you tell if an *uninterpretable* symbol system is writing? *Can* you tell?
  - If there are only single symbols (no texts); or symbols repeat too often; or symbols never repeat … then you'd probably guess it isn't
- But if there are "texts"; system seems to have structure?



What is writing and how do you tell if you've got it?

- But wait … what do you *mean* by writing?
  - Adapted from DeFrancis (1989, p. 4):
    - **Inclusivists**: Writing includes any system of graphic symbols that is used to convey some form of information.
    - **Exclusivists**: Writing includes only those systems of graphic symbols that encode *linguistic* entities (morphemes, syllables, phonemes …) and thus allow one to convey any information that can be conveyed in natural language.
  - An inclusivist definition from Powell (2009): Writing is "**a system of markings with a conventional reference that communicates information.**"

What is writing and how do you tell if you've got it?

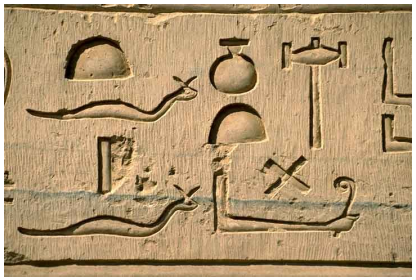
- OK, so can *statistical methods* tell you whether a system is writing or not?
- **First** be clear on inclusivism/exclusivism
  1. Per DeFrancis: be an inclusivist or an exclusivist – but *be consistent.*
  2. If you want to be an inclusivist then all you need to show (per Powell) was
     - System X was conventional
     - System X conveyed information
  3. … i.e. a very low bar, and usually not one that requires sophisticated statistical methods.
  4. Be aware that the person on the street is going to interpret **"**system X was writing**"** to mean that **system X was the same type of thing as what you are reading now.**
  5. Those who would make a substantive claim start by playing the exclusivist gambit.
- **Second:** structure does *not* imply that the system *encoded language.*

Babylonian deity symbols
(*kudurru* stones)

Mayan

Luwian

Egyptian



The Indus/Pictish Controversy 2004-2015

Indus “texts”: Indus Valley Civilization, 3rd Millennium BCE

Pictish symbols, 6th-9th century CE, Scotland

The Indus/Pictish Controversy 2004-2015

- **2004**: Farmer et al. argued that the cryptically short IVC texts were not writing (*exclusivist* sense).
- **2009**: Rao et al. in *Science* purport to provide "entropic evidence for linguistic structure" in the IVC symbols.
  - Various other papers followed (Rao et al. 2009; Rao, 2010)
- **2010**: Lee et al. *Proc. Roy. Soc.*, claim Pictish symbols "revealed as written language" via "Shannon entropy"
- Both were critiqued in the *Language Log* (here and here).
- **2010**: I published a critique in *Computational Linguistics*.
  - Replies by Rao et al (2010b) and Lee et al (2010b), and a reply by me (Sproat, 2010b)
- **2014**: My paper *Language* applied their methods + others to a larger set of non-linguistic symbols + writing.
  - The conclusion: none of the previously published methods are terribly useful.
  - Far simpler measures seem to be somewhat useful … but don't lead in the direction that Rao and Lee et al. want to go.
  - Again, replies by Rao et al (2014 – including Lee's team) and a reply by me (Sproat, 2015)
- Here, in the interests of time, I focus just on Lee et al.'s proposal, and its problems.
- **Note that the original Rao et al. and Lee et al. papers *still get cited* by people who believe their results**

.

Article

Restoring and attributing ancient texts using deep neural networks

https://doi.org/10.1038/s41586-022-04448-z

Received: 16 August 2021

Accepted: 19 January 2022

Published online: 9 March 2022

Open access

Check for updates

Yannis Assael[1,4✉], Thea Sommerschield[2,3,4✉], Brendan Shillingford[1], Mahyar Bordbar[1], John Pavlopoulos[4], Marita Chatzipanagiotou[4], Ion Androutsopoulos[4], Jonathan Prag[3] & Nando de Freitas[1]

Ancient history relies on disciplines such as epigraphy—the study of inscribed texts known as inscriptions—for evidence of the thought, language, society and history of past civilizations[1]. However, over the centuries, many inscriptions have been damaged to the point of illegibility, transported far from their original location and their date of writing is steeped in uncertainty. Here we present Ithaca, a deep neural network for the textual restoration, geographical attribution and chronological attribution of ancient Greek inscriptions. Ithaca is designed to assist and expand the historian's workflow. The architecture of Ithaca focuses on collaboration, decision support and interpretability. While Ithaca alone achieves 62% accuracy when restoring damaged texts, the use of Ithaca by historians improved their accuracy from 25% to 72%, confirming the synergistic effect of this research tool. Ithaca can attribute inscriptions to their original location with an accuracy of 71% and can date them to less than 30 years of their ground-truth ranges, redating key texts of Classical Athens and contributing to topical debates in ancient history. This research shows how models such as Ithaca can unlock the cooperative potential between artificial intelligence and historians, transformationally impacting the way that we study and write about one of the most important periods in human history.

Epigraphy is the study of texts—inscriptions—written directly on durable materials (stone, pottery, metal) by individuals, groups and institutions of the ancient world[2,3]. Thousands of inscriptions have survived to our time, but many have been damaged over the centuries and their texts are now fragmentary. Inscriptions may also be moved or trafficked far from their original location[4], and radiocarbon dating is unusable owing to the inorganic nature of most inscribed supports. Specialist epigraphers must then reconstruct the missing text, a process known as text restoration (Fig. 1), and establish the original place and date of writing, tasks known as geographical attribution and chronological attribution, respectively[5]. These three tasks are crucial steps towards placing an inscription both in history and within the world of the people who wrote and read it[6,7]. However, these tasks are non-trivial, and traditional methods in epigraphy involve highly complex, time-consuming and specialized workflows.

When restoring damaged inscriptions, epigraphers rely on accessing vast repositories of information to find textual and contextual parallels[8]. These repositories primarily consist of a researcher's mnemonic repertoire of parallels and, more recently, of digital corpora for performing 'string matching' searches. However, differences in the search query can exclude or obfuscate relevant results, and it is almost impossible to estimate the true probability distribution of possible restorations. Attributing an inscription is equally problematic—if it was moved, or if useful internal dating elements are missing, historians must find alternative criteria to attribute the place and date of writing (such as letterforms, dialects)[9]. Inevitably, a high level of generalization is often involved (chronological attribution intervals can be very long).

Deep learning for epigraphy

Here we overcome the constraints of current epigraphic methods by using state-of-the-art machine learning research. Inspired by biological neural networks, deep neural networks can discover and harness intricate statistical patterns in vast quantities of data[10]. Recent increases in computational power have enabled these models to tackle challenges of growing sophistication in many fields[11–14], including the study of ancient languages[15–18].

We present Ithaca, a deep neural network architecture trained to simultaneously perform the tasks of textual restoration, geographical attribution and chronological attribution. Ithaca, which was named after the Greek island that eluded the hero Odysseus' homecoming, was trained on inscriptions written in the ancient Greek language and across the ancient Mediterranean world between the seventh century BC and the fifth century AD. This choice was due to two main reasons. First, the variability of contents and context of the Greek epigraphic record, which makes it an excellent challenge for language processing; and second, the availability of digitized corpora for ancient Greek, an essential resource for training machine learning models.

[1]DeepMind, London, UK. [2]Department of Humanities, Ca' Foscari University of Venice, Venice, Italy. [3]Center for Hellenic Studies, Harvard University, Washington, DC, USA. [4]Department of Informatics, Athens University of Economics and Business, Athens, Greece. [5]Faculty of Classics, University of Oxford, Oxford, UK. [6]These authors contributed equally: Yannis Assael, Thea Sommerschield. ✉e-mail: yannisassael@google.com; thea.sommerschield@unive.it

Previous work

In recent years, several works have proposed traditional machine learning approaches to the study of ancient texts. This body of work has focused on optical character recognition and visual analysis[31–34], writer identification[35–37] and text analysis[38–44], stylometrics[45] and document dating[46]. It is only very recently that scholarship has begun to use deep learning and neural networks for optical character recognition[47–55], text analysis[56], machine translation of ancient texts[57–59], authorship attribution[60,61] and deciphering ancient languages[62,63], and been applied to study the form and style of epigraphic monuments[64].

40. Rao, R. P. et al. A Markov model of the Indus script. *Proc. Natl Acad. Sci. USA* **106**, 13685–13690 (2009).
41. Rao, R. P. et al. Entropic evidence for linguistic structure in the Indus script. *Science* **324**, 1165–1165 (2009).
42. Rao, R. P. et al. Entropy, the Indus script, and language: a reply to R. Sproat. *Comput. Linguist.* **36**, 795–805 (2010).

!!

[1]DeepMind, London, UK. [2]Department of Humanities, Ca' Foscari U

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Adamatzky A. 2022 Language of fungi derived from their electrical spiking activity. *R. Soc. Open Sci.* **9**: 211926. https://doi.org/10.1098/rsos.211926

Received: 18 December 2021
Accepted: 4 March 2022

Language of fungi derived from their electrical spiking activity

Andrew Adamatzky

Unconventional Computing Laboratory, UWE, Bristol, UK

AA, 0000-0003-1073-2662

Fungi exhibit oscillations of extracellular electrical potential recorded via differential electrodes inserted into a substrate colonized by mycelium or directly into sporocarps. We analysed electrical activity of ghost fungi (*Omphalotus nidiformis*), Enoki fungi (*Flammulina velutipes*), split gill fungi (*Schizophyllum commune*) and caterpillar fungi (*Cordyceps militaris*). The spiking characteristics are species specific: a spike duration varies from 1 to 21 h and an amplitude from 0.03 to 2.1 mV. We found that spikes are often clustered into trains. Assuming that spikes of electrical activity are used by

37. Lee R, Jonathan P, Ziman P. 2010 Pictish symbols revealed as a written language through application of Shannon entropy. *Proc. R. Soc. A* **466**, 2545–2560. (doi:10.1098/rspa.2010.0041)

Are the elaborate patterns of electrical activity used by fungi to communicate states of the mycelium and its environment and to transmit and process information in the mycelium networks? Is there a language of fungi? When interpreting fungal spiking patterns as a language, here we consider a number of linguistic phenomena as have been successfully used to decode pictish symbols revealed as a written language in [37]: (i) type of characters used to code, (ii) size of the character lexicon, (iii) grammar, (iv) syntax (word order), and (v) standardized spelling. These phenomena, apart from grammar and spelling, are analysed further.

The point here is *not* to argue about whether or not the Pictish/Indus symbols were writing.

Rather the point is to show that statistical methods are not as informative as people think…

and that a lot of the discussion has been based on fundamental misunderstandings.

Lee et al (2010)'s proposal

- Lee et al. claim their method is robust to "small sample size".
  - That's good, coz the entire Pictish corpus contains about 340 stones with about 650 total symbols (about 100 symbol *types*).
- Compute the *bigram conditional entropy* of a corpus

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \log p(y|x)$$

- Note that Rao et al. (2009) used bigram conditional entropy more or less directly.

- $$U_r = \frac{F_2}{\log_2(N_d/N_u)} \qquad C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d}$$

Bigram conditional entropy

Number of bigram types

Number of unigram types

Constant, derived as 7 by cross-validation

Total number of bigrams

Number of bigrams that occur once (*hapax legomena*)*

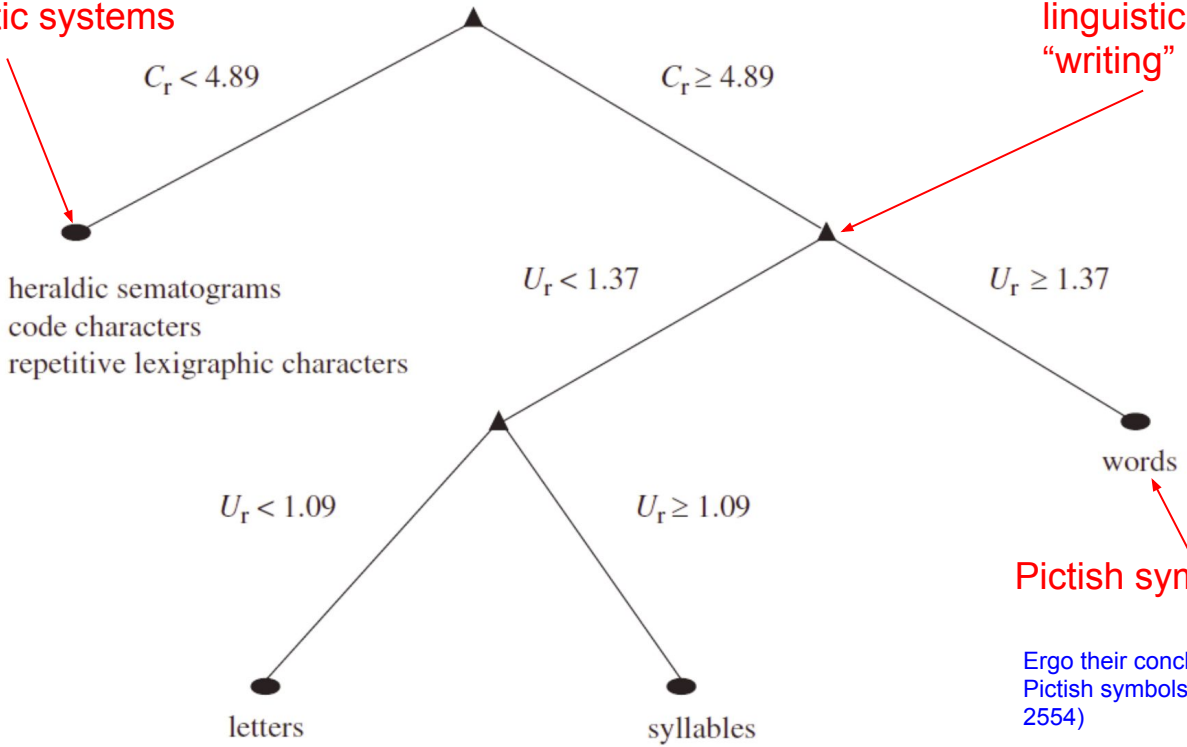
- Use U_r and C_r to train a decision tree
- Corpora: various writing systems, morse code, heraldry, random text

* This last term S_d/T_d will be familiar as the Good-Turing estimate of the probability mass of *unseen* events

Lee et al.'s decision tree

non-linguistic systems

linguistic systems:
"writing" (*exclusivist* sense)



Pictish symbols fall here

FIGURE 7. Reproduction of figure 6 from Lee et al. 2010a:9.

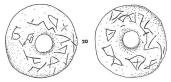
Ergo their conclusion that results suggest "that the Pictish symbols are lexigraphic* in nature" (page 2554)

(*i.e. what we'd normally call "logographic")

Sproat (2014)’s results using Lee et al’s approach

Corpus	Sample	Classification
Asian emoticons (21st cent)	|^_^| (◕‿◕)	
Pennsylvania barn stars (19th cent USA)		
Kudurrus (2nd millennium BCE Babylonia)		
Pictish symbols (1st millennium CE Scotland)		
Totem poles (19th cent Western North America)		
Vinča symbols (To 8 KyBP Danube region)		
Weather icon sequences (21st cent)	GMT 	
Various linguistic corpora	Amharic, Arabic, Chinese, Chinese Oracle Bones, Egyptian, English, Hindi, Korean (jamo and syllables) Linear B, Malayalam, Oriya, Sumerian, Tamil.	linguistic

Sproat (2014)'s results using Lee et al's approach

Corpus	Sample	Classification
Asian emoticons (21st cent)	|^_^| (◕‿◕)	linguistic: letters
Pennsylvania barn stars (19th cent USA)		linguistic: letters
Kudurrus (2nd millennium BCE Babylonia)		linguistic: syllables
Pictish symbols (1st millennium CE Scotland)		linguistic: words
Totem poles (19th cent Western North America)		linguistic: words
Vinča symbols (To 8 KyBP Danube region)		nonlinguistic
Weather icon sequences (21st cent)	GMT 	linguistic: letters
Various linguistic corpora	Amharic, Arabic, Chinese, Chinese Oracle Bones, Egyptian, English, Hindi, Korean (jamo and syllables) Linear B, Malayalam, Oriya, Sumerian, Tamil.	linguistic

Sproat (2014)'s results using Lee et al's approach

Corpus	Sample	Classification
Asian emoticons (21st cent)	|^_^| (◕‿◕)	linguistic: letters
Pennsylvania barn stars (19th cent USA)		linguistic: letters
Kudurrus (2nd millennium BCE Babylonia)		linguistic: syllables
Pictish symbols (1st millennium CE Scotland)		linguistic: words
Totem poles (19th cent Western North America)		linguistic: words
Vinča symbols (To 8 KyBP Danube region)		nonlinguistic
Weather icon sequences (21st cent)	GMT 	linguistic: letters
Various linguistic corpora	Amharic, Arabic, Chinese, Chinese Oracle Bones, Egyptian, English, Hindi, Korean (jamo and syllables) Linear B, Malayalam, Oriya, Sumerian, Tamil.	linguistic

Well, it got one right …

What feature worked best?

- In Sproat (2014) I looked at a bunch of features, including those of Rao et al. and Lee et al., and several others
- The most discriminative feature was a rather stupid measure:
  - *R* = number of tokens in a text that are repeats of a token previously in the text.
  - *r* = number of repeating tokens as in *R* that are *also adjacent to the token they repeat*
  - For AABACDB, *R* = 3 and *r* = 1
  - **Then compute *r/R***
- Interpretation: high *r/R* means the system has a lot of *local* repetition compared to the total amount of repetition.
- In linguistic terms, high *r/R* would mean, e.g., that when you have repetition, nearly all of it is morphological reduplication.

CORPUS	$\frac{r}{R}$
Barn stars	0.86
Weather icons	0.79
Sumerian	0.67
Totem poles	0.63
Vinča	0.59
Indus bar seals	0.58
Pictish	0.26
Asian emoticons	0.10
Egyptian	0.10
Mesopotamian deity symbols	0.099
Linear B	0.055
Oracle bones	0.048
Chinese	0.048
English	0.035
Arabic	0.032
Korean jamo	0.022
Malayalam	0.022
Korean	0.020
Amharic	0.018
Oriya	0.0075
Tamil	0.0060
Hindi	0.0017

The conclusions…

- Sproat (2014) also used a decision-tree classifier using a large set of features.
  - To cut a long story short: On most divisions of the data, it classified the Indus symbols **and** Pictish as **non-linguistic**
- Repetition measure *r/R,* and Lee et al's C_r both correlate with text length:
  - **Non-linguistic systems tend to have shorter texts!**
    - **That's a more important point than you might think.**
- OK, so I don't want to claim these results are definitive.
  - But if you want to take seriously the implicit claim of Rao & Lee et al. that statistical methods are useful, then you are bound to take seriously a result that points in the opposite direction of what you want to prove.

Lee et al.’s response & a timeline of position shifts

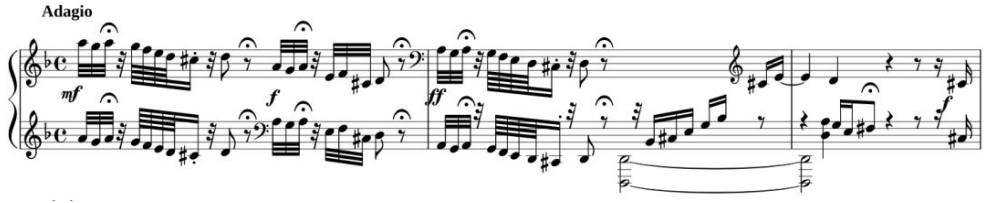
- Lee et al (2010a) clearly presented an ***exclusivist*** approach.
- In Sproat (2010a) I used Lee et al.’s method to show that *kudurru* symbols are “words”
  - Lee et al. (2010b) replicated this result and noted that this was a “difference in viewpoint over terminology as to the definition of what constitutes ‘writing’”, quoting Powell (2009).
  - In other words, *now* they are taking an ***inclusivist*** approach
- Using an updated version of the Babylonian deity symbol corpus, in Sproat (2014), *kudurru* symbols are now classified as “syllables” … Lee et al’s method is quite unstable.
- In their reply (Rao et al. 2014) they now claim that I misrepresented them, and that their method “was developed to try to determine the level of communication that a character communicates at, rather than a definition of writing” 😕
  - So, their original paper didn’t claim their method “revealed” Pictish symbols as “written language”? Hmm.
  - Also, why the change of position? Well for one thing in 2014 I had a lot more obviously non-linguistic systems and therefore a lot more silly results from their method.

So much for consistency … what about structure?

- E.g., Rao and colleagues frequently refer to the “rich syntactic structure” evidenced in the Indus texts.
- If “entropic” measures are useful at all, they should be able to distinguish structured symbol systems from ones that have no structure.
- Actually it’s not clear they do:
  - Liberman (2010, Language Log) showed that Lee et al.’s method classified 75 “texts” generated from successive tosses of seven cubic dice as a “syllabic writing system”
- But suppose they *did* reveal structure? What would that tell us?

Structure ≠ Linguistic Structure

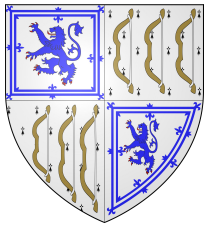
- For example



$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} \right)$$

Structure ≠ Linguistic Structure

- One of my favorite examples: (British) heraldry vs. Japanese *kamon* (家紋)
- Lots of similarities:
  - Both were used to represent either clans (*kamon*) or (in British heraldry at least) individuals.
  - Both were used in battle to identify armies; they had to be easily identifiable from afar.
  - Both frequently made use of stylized depictions of plants and animals.
  - Both also made use of many simple geometrical figures.
  - Written language could be incorporated into the design in both cases.
  - In both it was common for the motif to allude to some property of the family name.

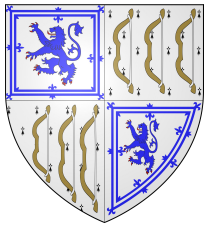


Bowes-Lyon family (source: Wikipedia)
– an example of *canting* arms (*armes parlantes)*

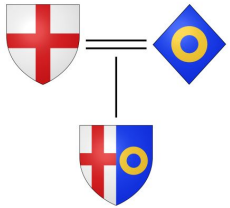
“Falling wisteria” 下り藤 (*sagari fuji*) *kamon* of the 藤井 *Fujii* (“wisteria well”) family, Numata, Gunma Prefecture.

Structure ≠ Linguistic Structure

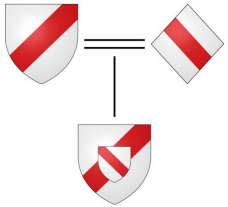
- The *Bowes-Lyon* case illustrates a property of heraldry that is *not* found in *kamon*: quartering
  - Quartering occurred when a wife, an heiress with children, dies. Upon her death, the son of the marriage may quarter his mother's arms with those of his father
  - Marriage could also be represented by *impalement*
    - If the wife is *not* a heraldic heiress, the husband *impales* her arms on the *sinister* (left) side of his arms
  - … or if she *is* a heraldic heiress, her arms are placed over his on an *escutcheon of pretense*



Quartering



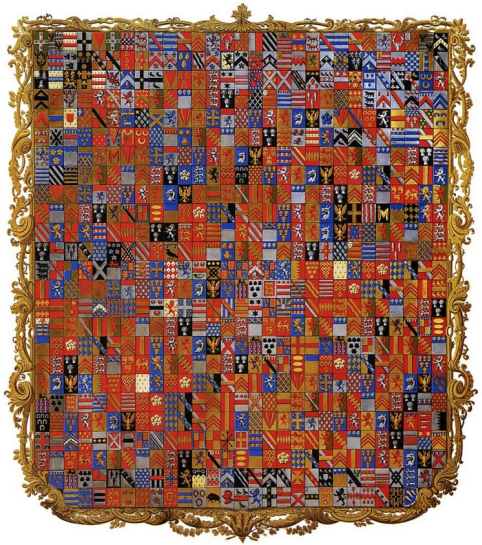
Impalement (source: Wikipedia)



Escutcheon of pretense (source: Wikipedia)

Structure ≠ Linguistic Structure

- Quartering in particular theoretically allowed for unbounded depth



The arms of George, Marquess of Buckingham

The arms of Thomas Stanley, Earl of Derby (d. 1572) – a more typical example of complex quartering

Structure ≠ Linguistic Structure

- **There is no equivalent of any of these devices in *kamon***
  - In marriage, only the male line was considered important (Tonomura, 1990)
  - Cf. 婿養子*mukoyōshi*, whereby a family without a male heir could adopt an adult as their son
- Note that there were female *mon*, called *onnamon* 女紋 (Morimoto, 2006).
  - These were *only* inherited in the female line (typically mother-to-daughter)
  - They were *never* incorporated into the husband's *mon*
- Two systems that are essentially similar in their basic function, can diverge *in their syntactic complexity* due to simple differences in cultural constraints on their use.

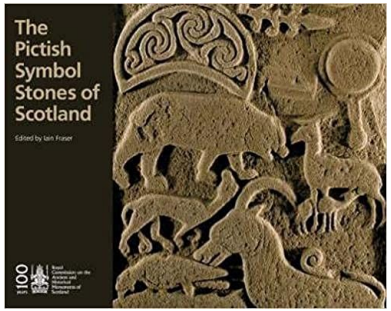
But nobody’d think heraldry is written language, right?

- If nothing else, the syntax is 2D
- In writing systems, there may be 2D arrangements of symbols *locally*, but writing is generally *linear* macroscopically.
- So nobody would expect something with apparently 2D arrangements of symbols to be *revealed as written language*.

But nobody’d think heraldry is written language, right?

- If nothing else, the syntax is 2D
- In writing systems, there may be 2D arrangements of symbols *locally*, but writing is generally *linear* macroscopically.
- So nobody would expect something with apparently 2D arrangements of symbols to be *revealed as written language*.

World History Encyclopedia



Archaeotravel

Pictish symbols only look “linear” in electronic encodings

Structure ≠ Linguistic Structure

- Structure in a system derives from what the system is used to represent.
- Structure implies *nothing* about whether a system represents language or not.
- So the real question is not whether your method can detect structure…
  - … but whether the method can specifically detect *linguistic* structure.

- So far I have not seen such a method.

Summary

- Statistical methods are a double-edged sword:
  - You don’t get to pick just the methods that show what you want.
- Above all, one has to be consistent in how one uses terms:
  - Immense amounts of confusion caused by fuzzy ideas about what the term “writing” means.
- Structure in a symbol system relates only to one thing:
  - Whether the information you are trying to encode itself has structure
  - **That should have been obvious from the get go…**
- Results of computational approaches – so far – have been mostly negative:
  - But they have laid bare important inconsistencies and fallacies.

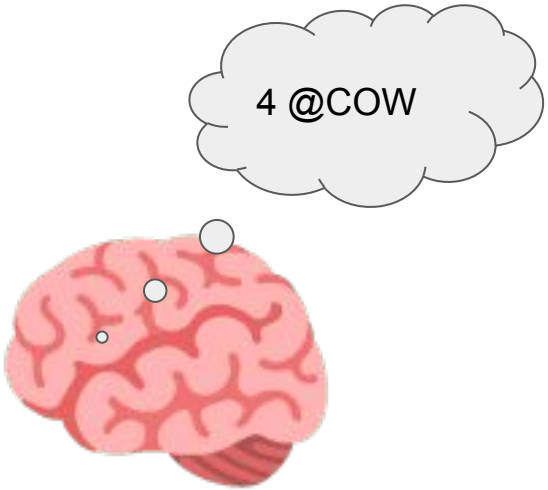
Evolution of Writing

(synopsis only)

Modeling the evolution of writing *ex nihilo*

- Sometimes computational methods can help where hard evidence is scarce.
  - How did *non-linguistic* symbol systems evolve into writing?
  - How is the evolution affected by the *shape of the language's morphemes*?
  - Sproat (2017), Sproat (forthcoming)
- Basic idea:
  - Generate artificial "languages" with differently shaped morphemes:
    - Monosyllabic; (maximally) disyllabic; (maximally) sesquisyllabic
    - "Morpheme" is just an association between *phonological form* and *concept*
    - Randomly associate a small (100) number of *concepts* with symbols
  - Generate a set of "texts" in each language pairing
    - "meaning" → symbol strings
    - "meaning" + "phonetics" → symbol strings
  - Train a neural model to produce symbols under these two conditions (Sproat, forthcoming)
  - Observe how the model extends to new **meaning** or **meaning+phonetics** "messages"
  - Iterate the process and see how the system evolves over time:
    - What proportion of *new* **meanings** or **meaning+phonetics** acquire written forms?

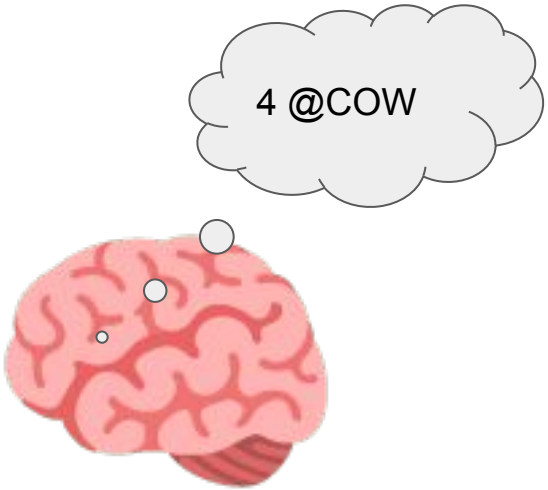
Training scenario



4 @COW		IV 🐄
3 @HORSE		III 🐎
7 @PIG		VII 🐖

- Present the meaning alone
  - Simulates an accountant just using the symbols for their meaning.
- **Task: write the correct symbols…**

Training scenario

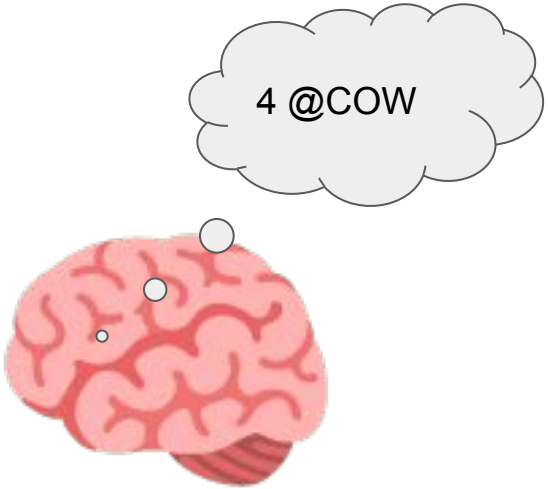


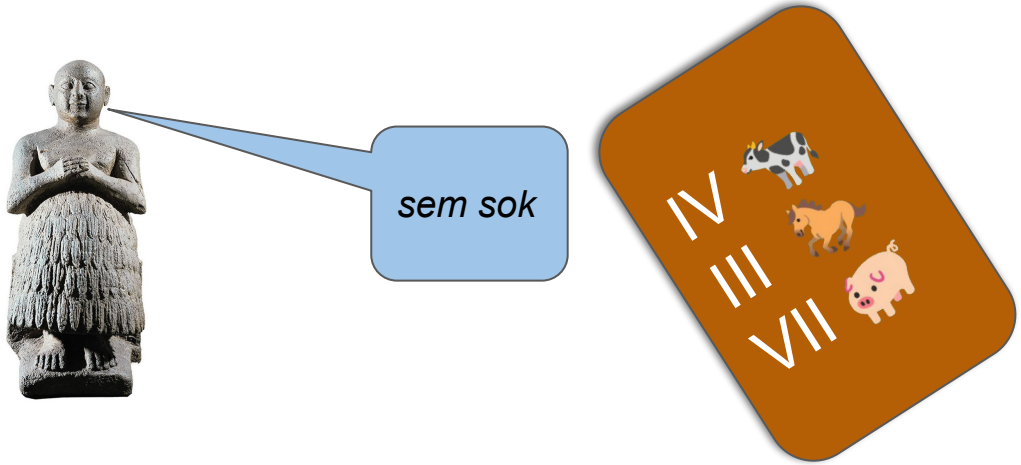
Hypothesis: Writing evolved in an institutional context in which symbols were effectively *dictated*, so that the user of the symbol system gradually came to associate the symbols with sounds.

4 @COW		IV 🐄
3 @HORSE		III 🐎
7 @PIG		VII 🐖

- Present the meaning alone
  - Simulates an accountant just using the symbols for their meaning.
- **Task: write the correct symbols…**

Training scenario

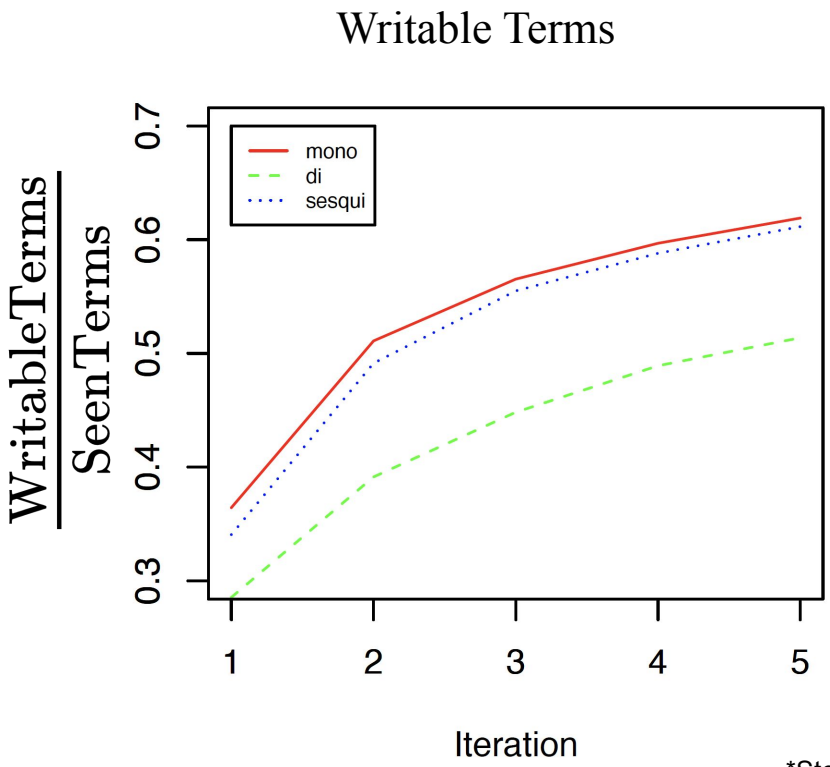




4 @COW	*sem sok*	IV 🐄
3 @HORSE	*ka bin*	III 🐎
7 @PIG	*nan kom*	VII 🐖

- … also present the phonology
  - Simulates someone “dictating” the account to the accountant.
- **Task: write the correct symbols…**

Easier generalization in “monosyllabic languages”*

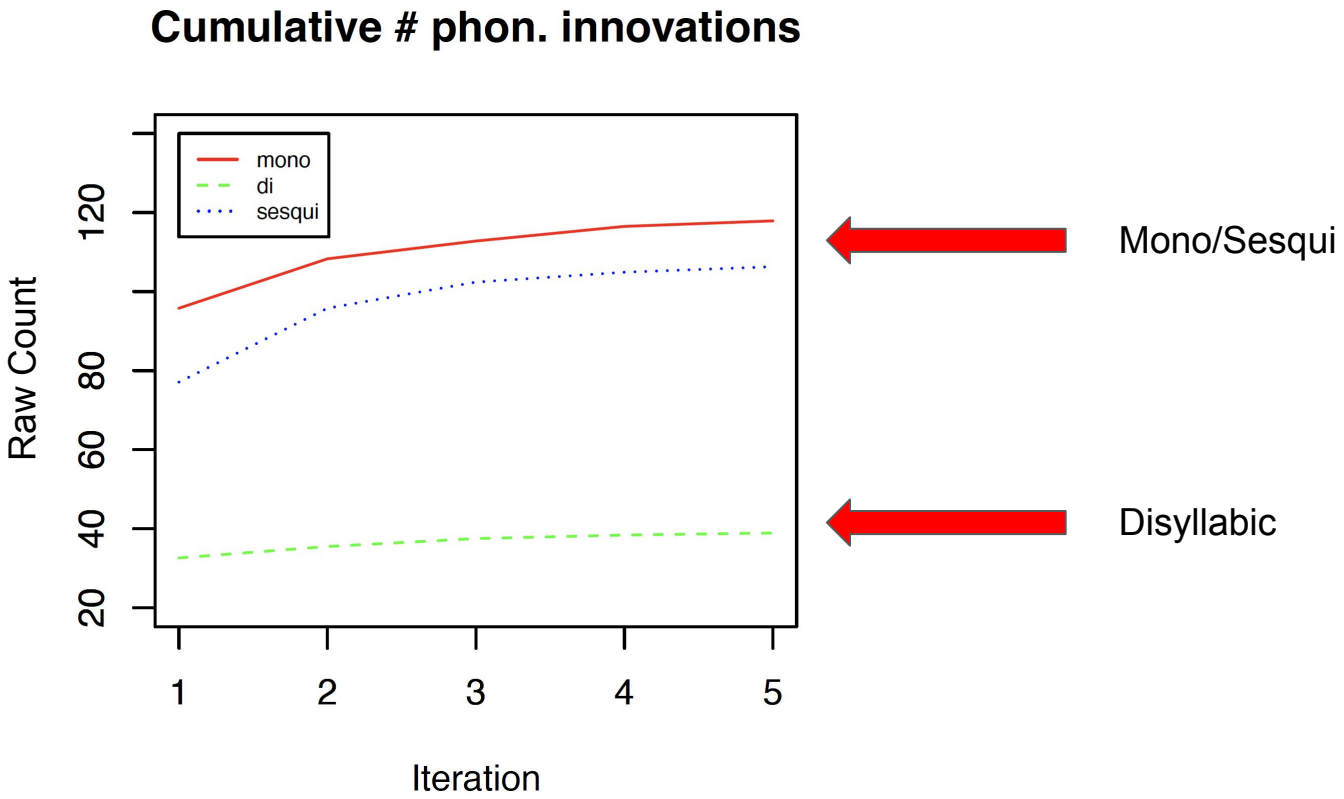


Mono/Sesqui

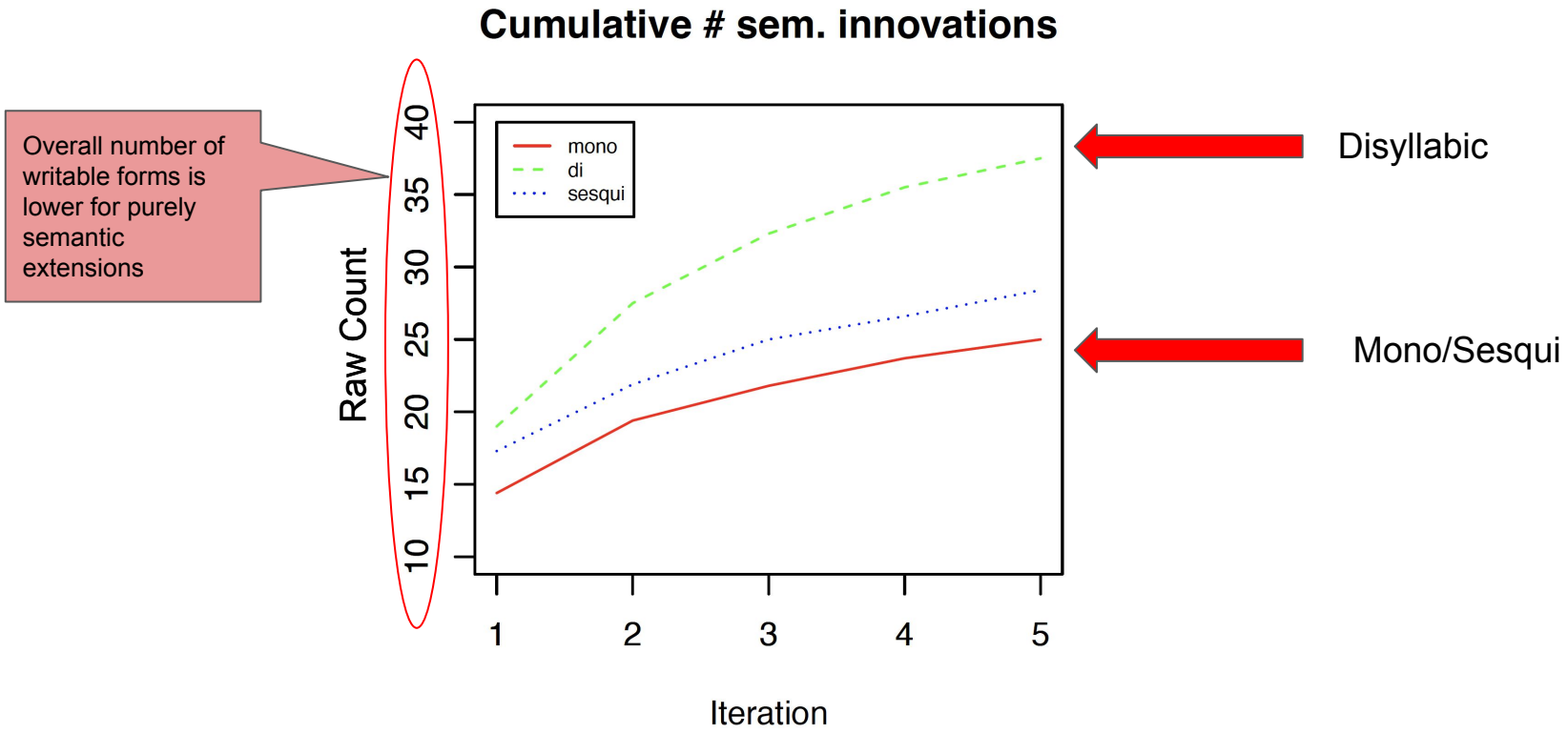
Disyllabic

*Steinthal (1852), Daniels (1992), Boltz (2000), Buckley (2008)

Fewer phonological extensions with disyllabic languages



More semantic extensions with disyllables:
but semantic extension is less “efficient” overall.



Summary

- Simulation mimics some things we know about the early evolution of writing:
  - Symbols could be extended in use on the basis of their meaning
  - Symbols could be extended in use on the basis of their sound
- **Phonology is a more productive means of extension than semantics.**
- Phonological properties of the language are important:
  - It is easier to generalize in a monosyllabic/sesquisyllabic language than a disyllabic language
    - It’s harder to find a close neighbor for disyllables than monosyllables/sesquisyllables.

Further Thoughts

Further thoughts: A few future directions

- Need more work on understanding differences among writing systems.
- A better understanding of the mathematical properties of symbol systems:
  - Is it really possible to determine a system's function *just* by looking at symbol distributions?
  - Can one distinguish *structure* from specifically *linguistic structure*?
- More sophisticated simulations of the evolution of writing:
  - Using real glyph images and real speech…
- **Automated "decipherment"?** Cf. recent papers by Barzilay's group at MIT
  - E.g.: Luo et al. (2019, 2020) – see my critique of the latter paper here.
  - I'm skeptical that this will prove useful:
    - Mostly post-hoc decipherments of systems we already know: Ugaritic, Linear B…
    - Unrealistic expectations about finding related languages; complexity of writing system.
      - Segmental script for a language w/ closely related known language ≫ easier than …
      - A mixed semantic-syllabic script for an unknown language.
    - And there is still the problem of small sample sizes…

References

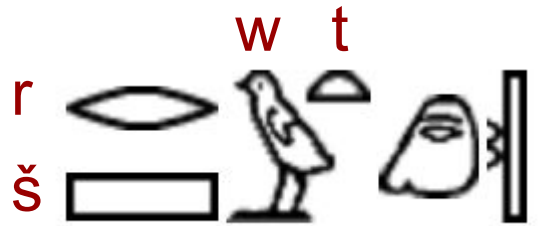
References are available here.

Common properties of early writing systems

- All evolved methods for encoding *phonology*
- All were *mixed* with *semasiographic* and *phonographic* symbols: "semantic-phonetic" constructions

鯉 *lǐ* 'carp' < 魚FISH + 里*lǐ*

ba  BALAM 'jaguar'

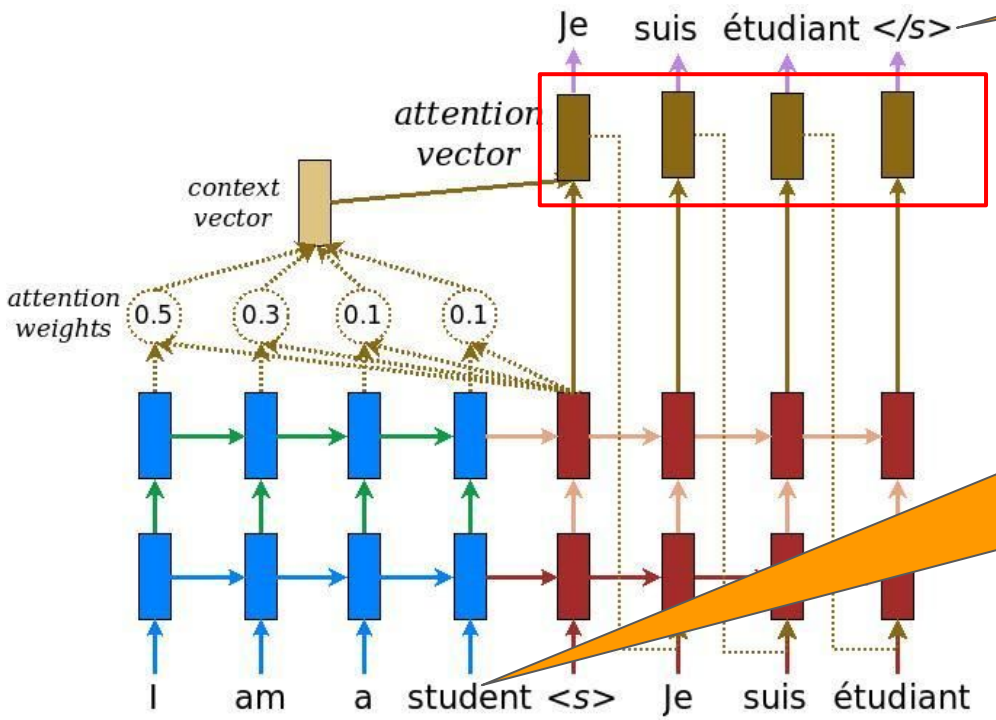


ršwt 'joy'


ma

$urim_5$ + CITY = "Ur"

Attention based model



In our case: output is symbols

In our case the input is *embeddings* of

- Meaning
- Meaning + Phonology

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

Embeddings

- Symbols are associated with meanings.
  - “Meaning” here is implemented using *word embeddings* as a proxy (in this case from the British National Corpus – Fares et al. 2017)
  - Two similar meanings are close in embedding space.
- If the system is also trained with “dictation”, symbols are presented along with *phonetic embeddings*
  - E.g. two similar syllables are close in embedding space.

Phonetic embeddings

- Phonetic distance:
  - Rhyme distances
    - Exact rhymes: ***et ~ et***
    - Close rhymes (sharing place/manner of final C): ***et ~ ed***
    - Non-rhyme: ***et ~ en***
  - Onset distances:
    - Exact alliteration: ***t ~ t***
    - Close alliteration: ***t ~ d***
    - Non-alliteration: ***t ~ n***
- This approximates, e.g., the case in Chinese (Baxter 1992, p. 348)*
- 300-length vectors are then created for each syllable in the language and its phonetic distance to the 300 most frequent syllables.
  - Similar approach is taken for disyllables.

"to be written with the same phonetic element, words must normally have identical main vowels and codas, and their initial consonants must have the same position of articulation"

Phonetic embeddings

- Phonetic distance:
  - Rhyme distances
    - Exact rhymes: ***et ~ et***
    - Close rhymes (sharing place/manner of final C): ***et ~ ed***
    - Non-rhyme: ***et ~ en***
  - Onset distances:
    - Exact alliteration: ***t ~ t***
    - Close alliteration: ***t ~ d***
    - Non-alliteration: ***t ~ n***
- This approximates, e.g., the case in Chinese (Baxter 1992, p. 348)*
- 300-length vectors are then created for each syllable in the language and its phonetic distance to the 300 most frequent syllables.
  - Similar approach is taken for disyllables.

Most similar	Least similar
bal ↔ fal	ol ↔ sur
bar ↔ far	ur ↔ fol
bay ↔ fay	dur ↔ kol
bel ↔ fel	ur ↔ pol
ber ↔ fer	ol ↔ dur
bey ↔ fey	ur ↔ xol
bop ↔ fop	ur ↔ dol
buy ↔ fuy	ur ↔ sol
fet ↔ bet	ur ↔ kol
fok ↔ bok	ur ↔ tol

Most and least similar syllables

"to be written with the same phonetic element, words must normally have identical main vowels and codas, and their initial consonants must have the same position of articulation"

Phonetic embeddings

- Phonetic distance:
  - Rhyme distances
    - Exact rhymes: ***et ~ et***
    - Close rhymes (sharing place/manner of final C): ***et ~***
    - Non-rhyme: ***et ~ en***
  - Onset distances:
    - Exact alliteration: ***t ~ t***
    - Close alliteration: ***t ~ d***
    - Non-alliteration: ***t ~ n***
- This approximates, e.g., the case in Chinese (Baxter 1992, p. 348)*
- 300-length vectors are then created for each syllable in the language and its phonetic distance to the 300 most frequent syllables.
  - Similar approach is taken for disyllables.

Most similar	Least similar
mun.mun ↔ mun	ok.bem ↔ yak.ok
fek.tor ↔ fek.dor	iy.gem ↔ gun.fiy
suk.du ↔ suk.tu	iy.gem ↔ yok.un
fey.kuy ↔ fey.xuy	ok.sok ↔ fak.iy
gum.ko ↔ gum.xo	iy.gem ↔ xem.dak
kom.xin ↔ xom.xin	sak.iy ↔ iy.xok
kor.kok ↔ xor.kok	iy.gem ↔ gem.yum
rup.xek ↔ rup.kek	iy.gem ↔ yak.ok
rur.xiw ↔ rur.kiw	iy.gem ↔ fak.iy
tal.kum ↔ tal.xum	gok.ak ↔ iy.xok

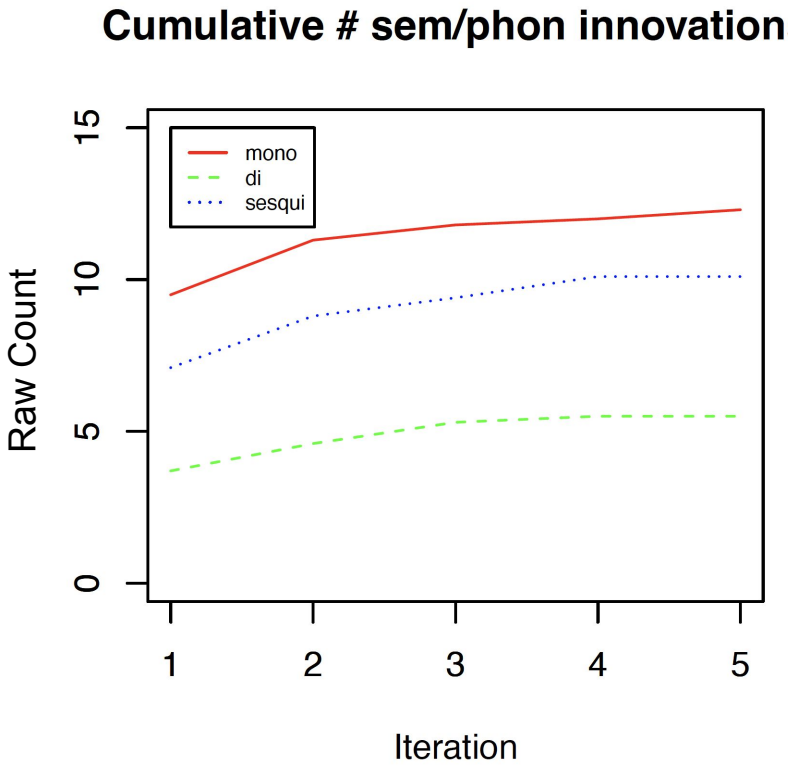
Most and least similar disyllables

"to be written with the same phonetic element, words must normally have identical main vowels and codas, and their initial consonants must have the same position of articulation"

Sketch of training cycle

1. Train the system to learn meaning/meaning+phonology → symbols
2. See what the system does with terms (meaning/meaning+phonology) it has not been trained on.
    - The system will always predict *some* output for a novel input, but how *confident* is it?
    - Confidence = difference between score for first and second predictions.
3. If confidence is above a certain threshold:
    - For semantic: extend to terms with similar meanings
    - For phonetic: extend to terms with similar sounds
    - If both above threshold, then model creates a semantic-phonetic compound
4. Update the set of writable terms, and generate a new set of training texts.
5. Go to 1. and repeat.
6. Look at the evolution of the writable terms relative to the total number of seen terms.

Fewer sem+phon extensions in disyllabic languages

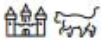
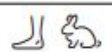
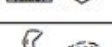
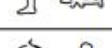


Mono/Sesqui

Disyllabic

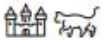
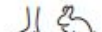
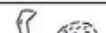
Some semantic/phonetic compounds from one run

Novel term with this meaning and sound

Sem.	Pron.	Proposed Form	Glyph Sem./Glyph Phon.
@GATEHOUSE	gan		@CASTLE/gan
@CARP	kuy		@FISH/xuy
@INCH	fet		@FOOT/pet
@PONY	far		@HORSE/far
@MOTEL	yiw		@HOTEL/tiw
@THIGH	op		@LEG/op
@PARSLEY	kol		@ONION/xol

Some semantic/phonetic compounds from one run

Novel term with this meaning and sound

Sem.	Pron.	Proposed Form	Glyph Sem./Glyph Phon.
@GATEHOUSE	gan		@CASTLE/gan
@CARP	kuy		@FISH/xuy
@INCH	fet		@FOOT/pet
@PONY	far		@HORSE/far
@MOTEL	yiw		@HOTEL/tiw
@THIGH	op		@LEG/op
@PARSLEY	kol		@ONION/xol

@CARP *lǐ* 鯉 @FISH(魚)/*lǐ*(里) (= ‘village’)