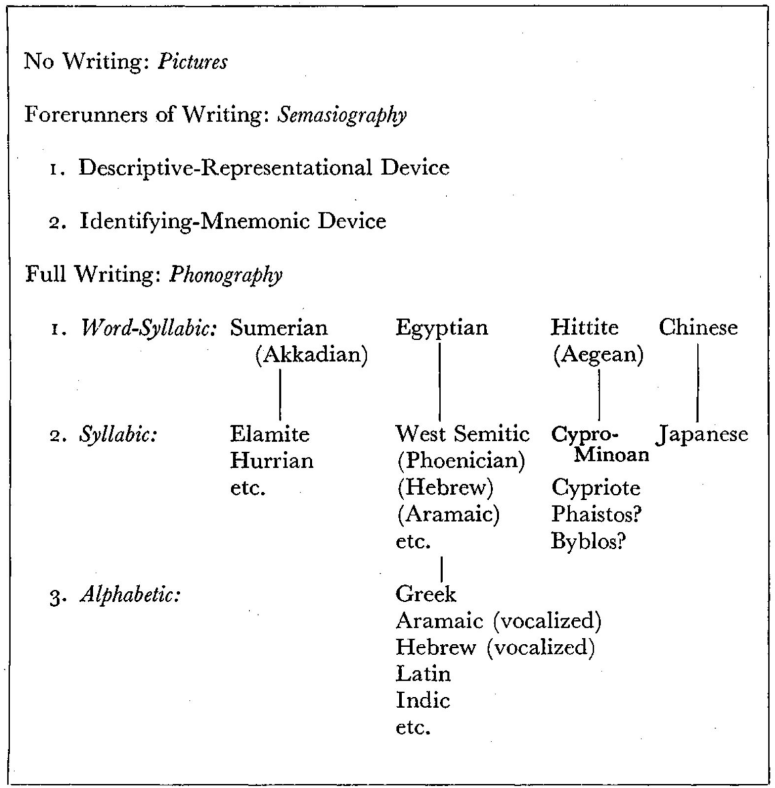
Taxonomy of Writing Systems: How to measure how logographic a system is

Richard Sproat & Alexander Gutkin
Google, Japan & Google, UK
rws@google.com, agutkin@google.com
SIGTYP 2020

Background: taxonomies of writing systems



No Writing: *Pictures*

Forerunners of Writing: *Semasiography*

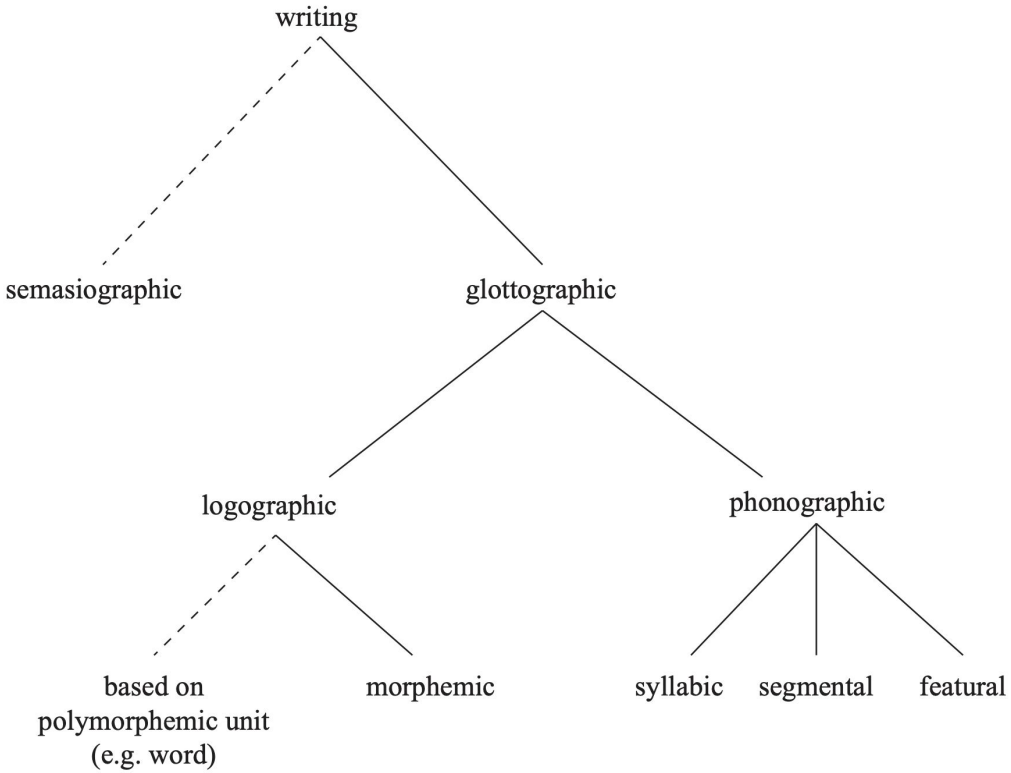
1. Descriptive-Representational Device
2. Identifying-Mnemonic Device

Full Writing: *Phonography*

1. *Word-Syllabic:*	Sumerian (Akkadian)	Egyptian	Hittite (Aegean)	Chinese
2. *Syllabic:*	Elamite Hurrian etc.	West Semitic (Phoenician) (Hebrew) (Aramaic) etc.	Cypro-Minoan Cypriote Phaistos? Byblos?	Japanese
3. *Alphabetic:*		Greek Aramaic (vocalized) Hebrew (vocalized) Latin Indic etc.		

Gelb 1952

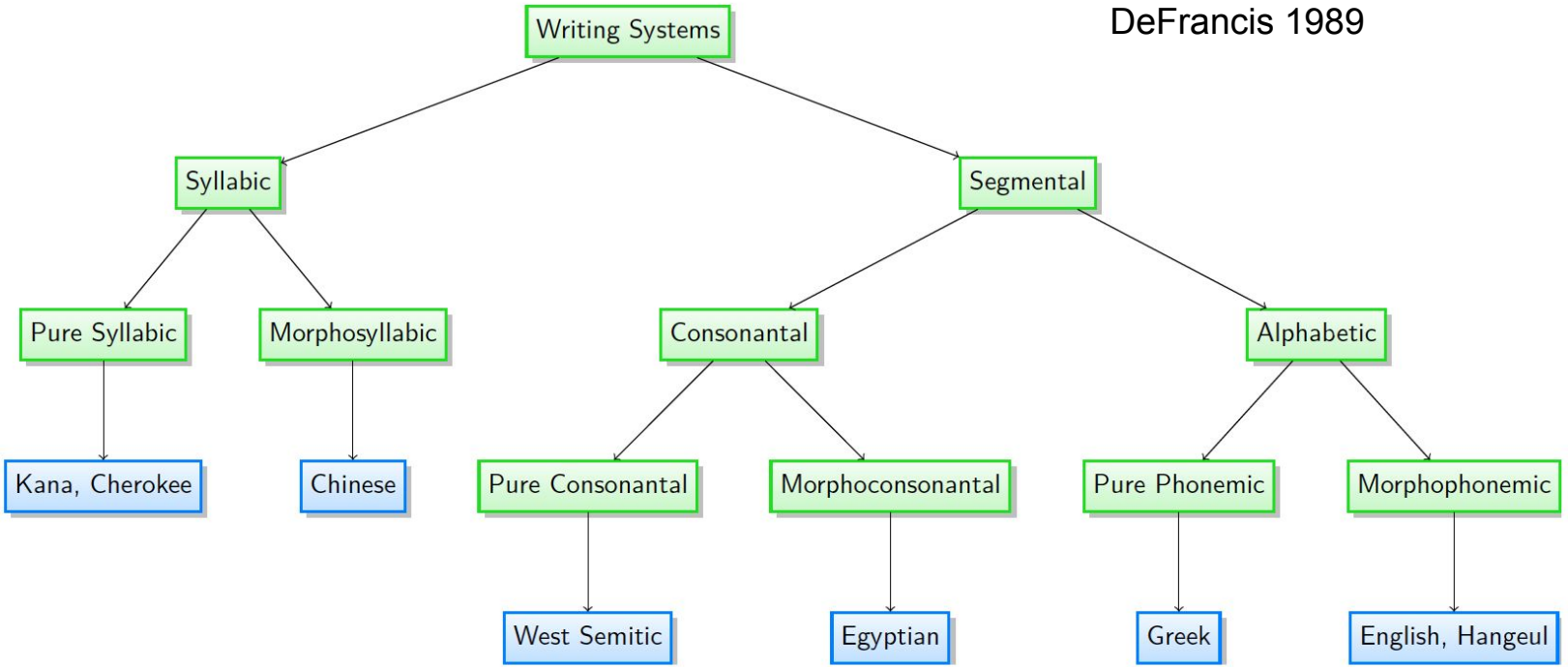
Background: taxonomies of writing systems



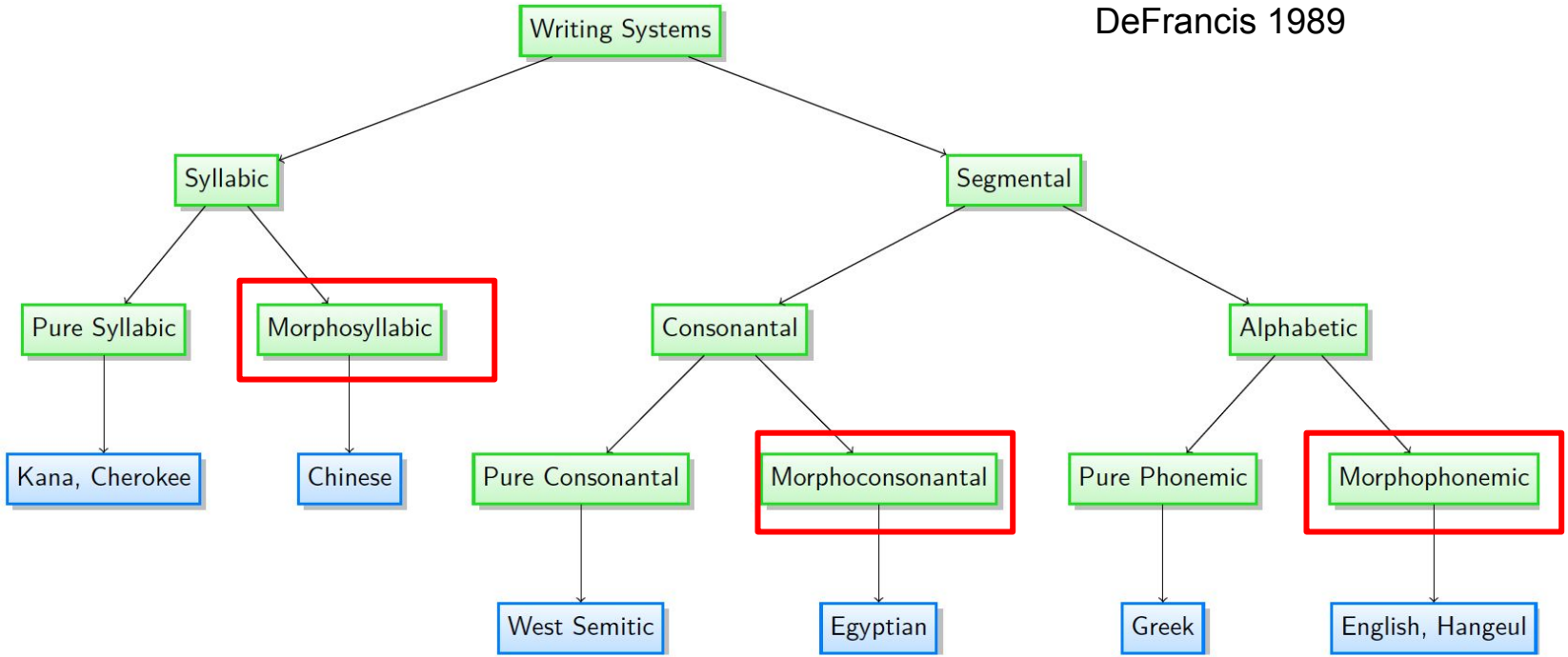
Sampson 1985

Background: taxonomies of writing systems

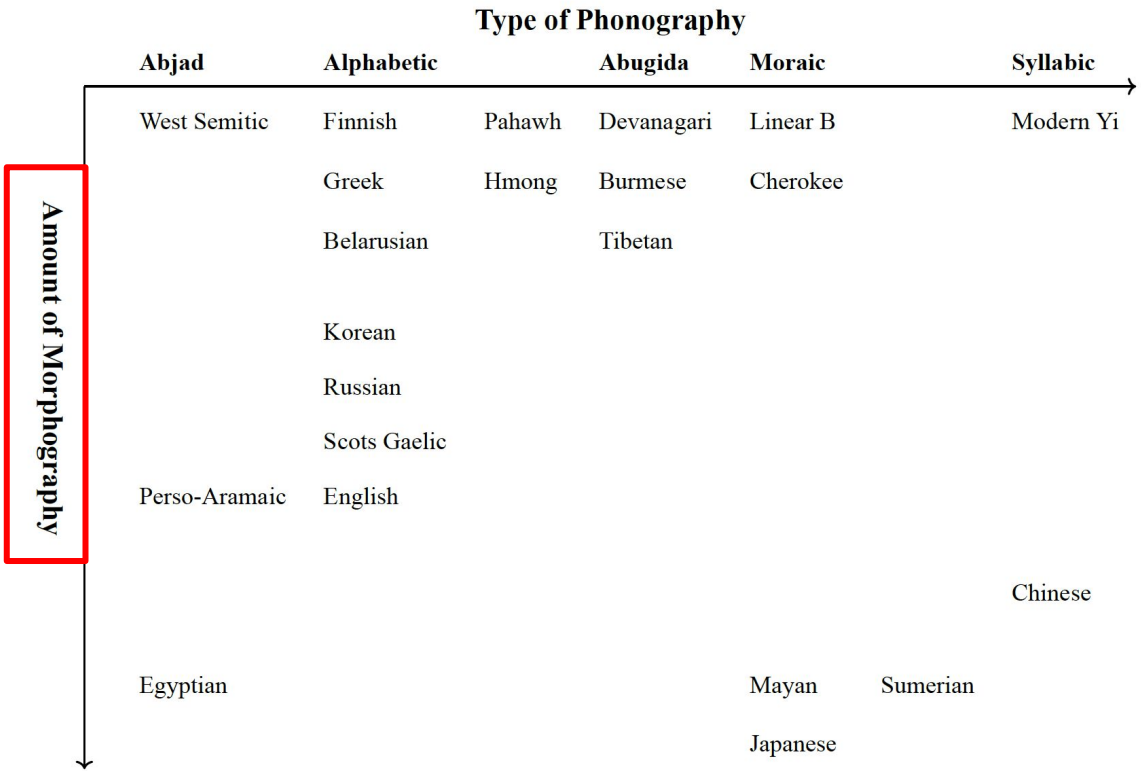
DeFrancis 1989



Background: taxonomies of writing systems



Background: taxonomies of writing systems



Rogers (2005) (< Sproat 2000)

What *is* logography?

- The term is actually never really defined in the literature
- But it seems to relate to two ideas:
  - Different words/morphemes should be spelled differently even if pronounced the same:
    - **distinct homophones**
    - Cf: *pear, pair, pare*
  - The same morpheme should be uniformly spelled despite morphophonological changes:
    - **uniform spelling**
    - Cf: *telegraph* /ˈtɛləgɹæf/, *telegraphy* /təˈlɛgɹəf-/, *telegraphic* /tɛləˈgɹæf-/
- We concentrate here on the **distinct homophones** notion, leaving **uniform spelling** for future research

Outline

- Taxonomies of writing systems
- What is “logography”?
- A previous computational proposal for measuring the degree of logography
- Three computational measures of logography:
  - A simple lexical measure
  - An entropic measure
  - An attention-based measure
- Experiments and results
- Some conclusions

How do you measure logography?

- One proposal by Penn & Choma (2006): correlation coefficients
- Basic idea: because they represent words/morphemes so, indirectly, meaning
  - logographic symbols should be more “bursty” in their cooccurrence within a document
  - conversely, phonographic symbols should cooccur more uniformly

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma(X)\sigma(Y)},$$

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

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

“each grapheme type is treated as a variable, and each document represents an observation”

cov(X, Y) is the covariance between *X* and *Y*

μ_i is the mean of the *i*th grapheme

Penn & Choma's experiments

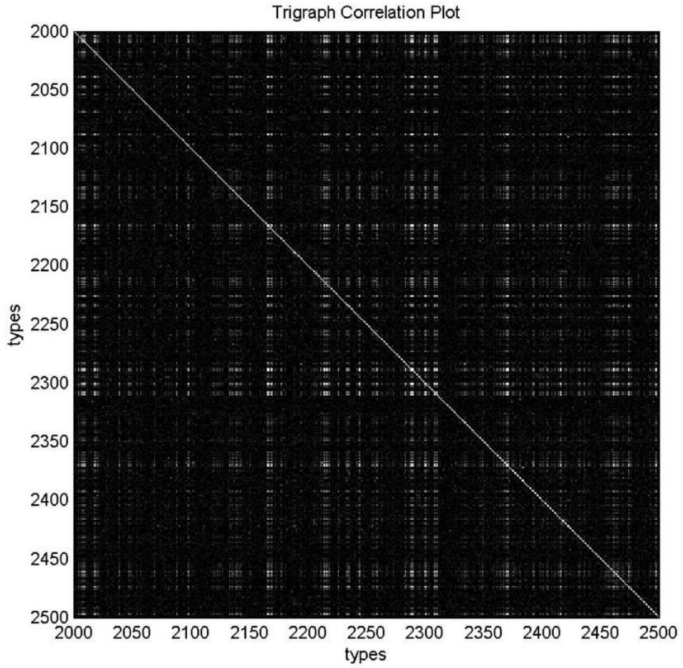
- Compared Chinese and "trigrammed" English:
  - The point was to find a nominally phonographic system that has roughly the order of magnitude of the number of Chinese characters
  - Penn & Choma would have preferred to use Yi, a syllabic system with a large number of symbols
- Corpora were:
  - A "Chinese news corpus"
  - The Brown corpus

Penn & Choma's experiments

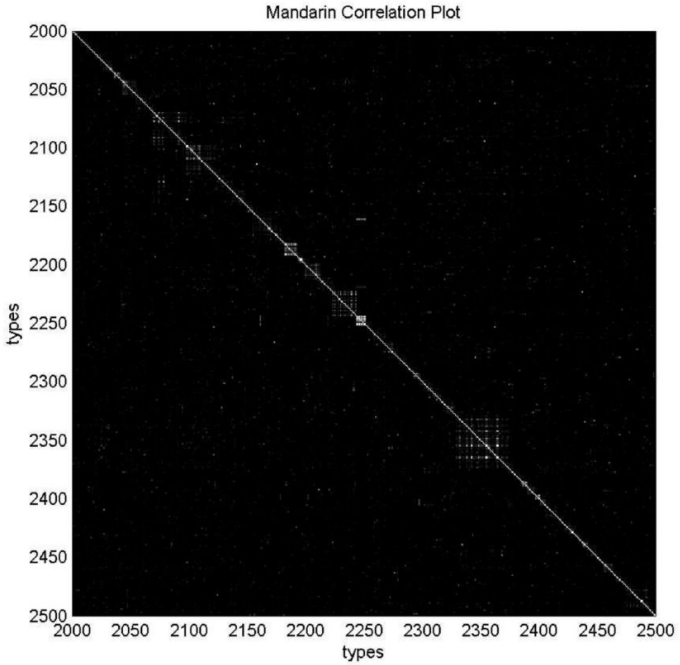
- Compared Chinese and "trigrammed" English:
  - The point was to find a nominally phonographic system that has roughly the order of magnitude of the number of Chinese characters
  - Penn & Choma would have preferred to use Yi, a syllabic system with a large number of symbols
- Corpora were:
  - A "Chinese news corpus"
  - The Brown corpus

Source: https://en.wikipedia.org/wiki/Nuosu_language#/media/File:Yi_words_sign.jpg

Penn & Choma (2006)



(a) English

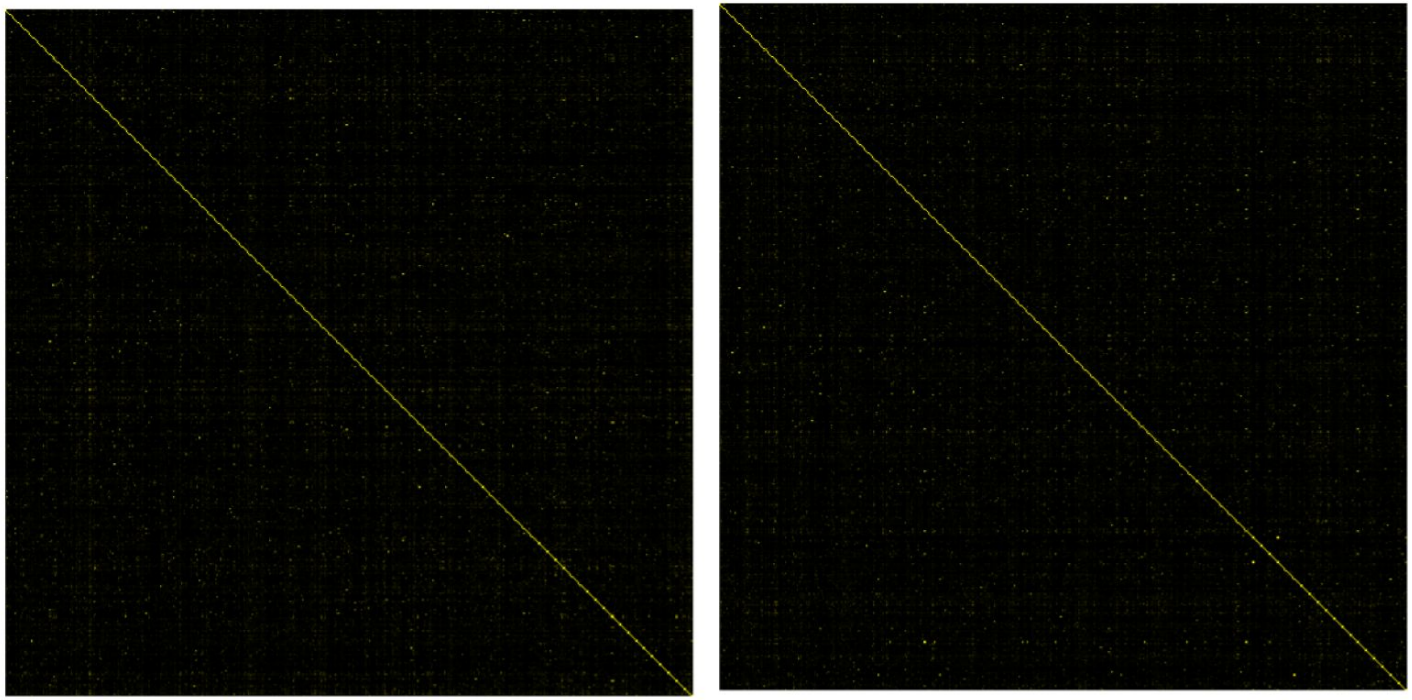


(b) Chinese

Penn & Choma: Problems

- We attempted to replicate Penn & Choma's result using the Bible Corpus (Christodoulopoulos & Steedman, 2015), taking each chapter as a document
- This fails,

Penn & Choma (2006): problems



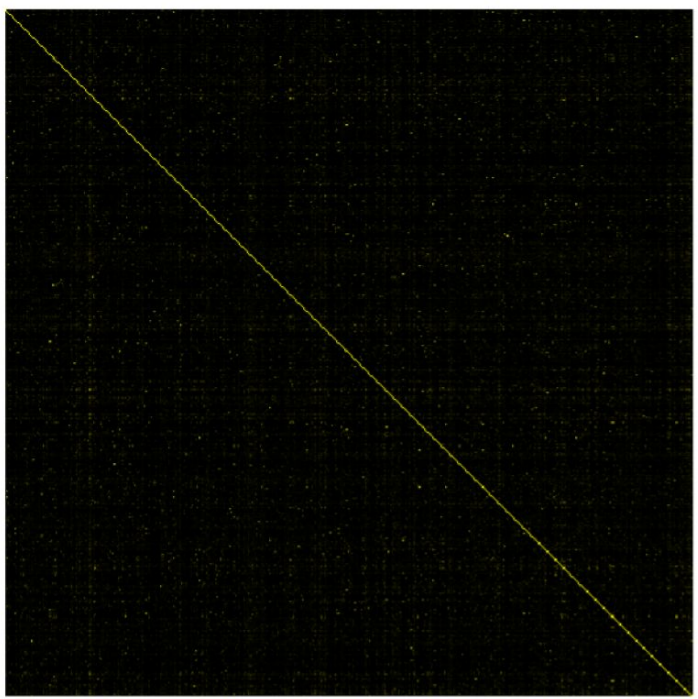
(a) English

(b) Chinese

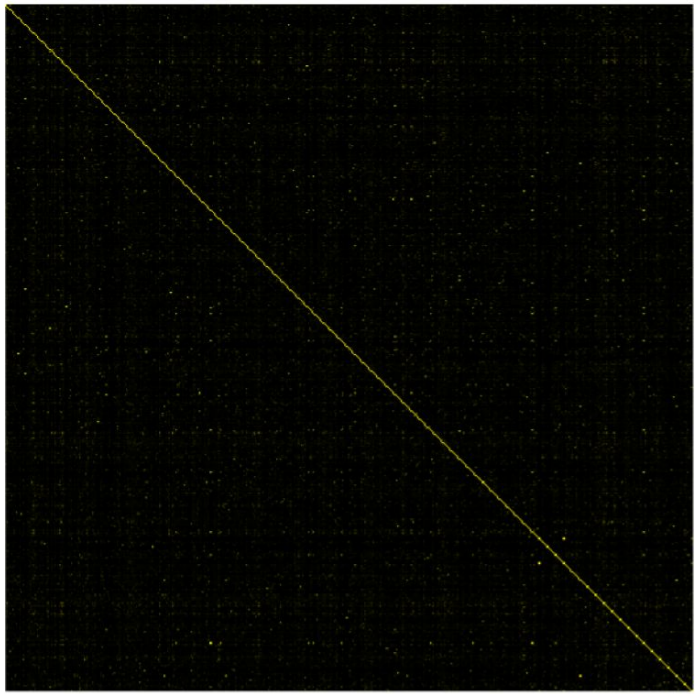
Penn & Choma: Problems

- We attempted to replicate Penn & Choma's result using the Bible Corpus (Christodoulopoulos & Steedman, 2015), taking each chapter as a document
- This fails, and the reason seems to be because of the document sizes:
  - Brown corpus: 2000 words per document, i.e. about 4000 trigrams
  - Chinese news corpus (e.g. Chinese Gigaword) has about 450 characters per document
  - Bible:
    - Approx. 1100 trigram letters per document (chapter) for English
    - Approx. 780 characters per document for Chinese
    - Group 6 chapters into a "document" for English: about 6600 trigram letters per document

Replication with larger documents for English

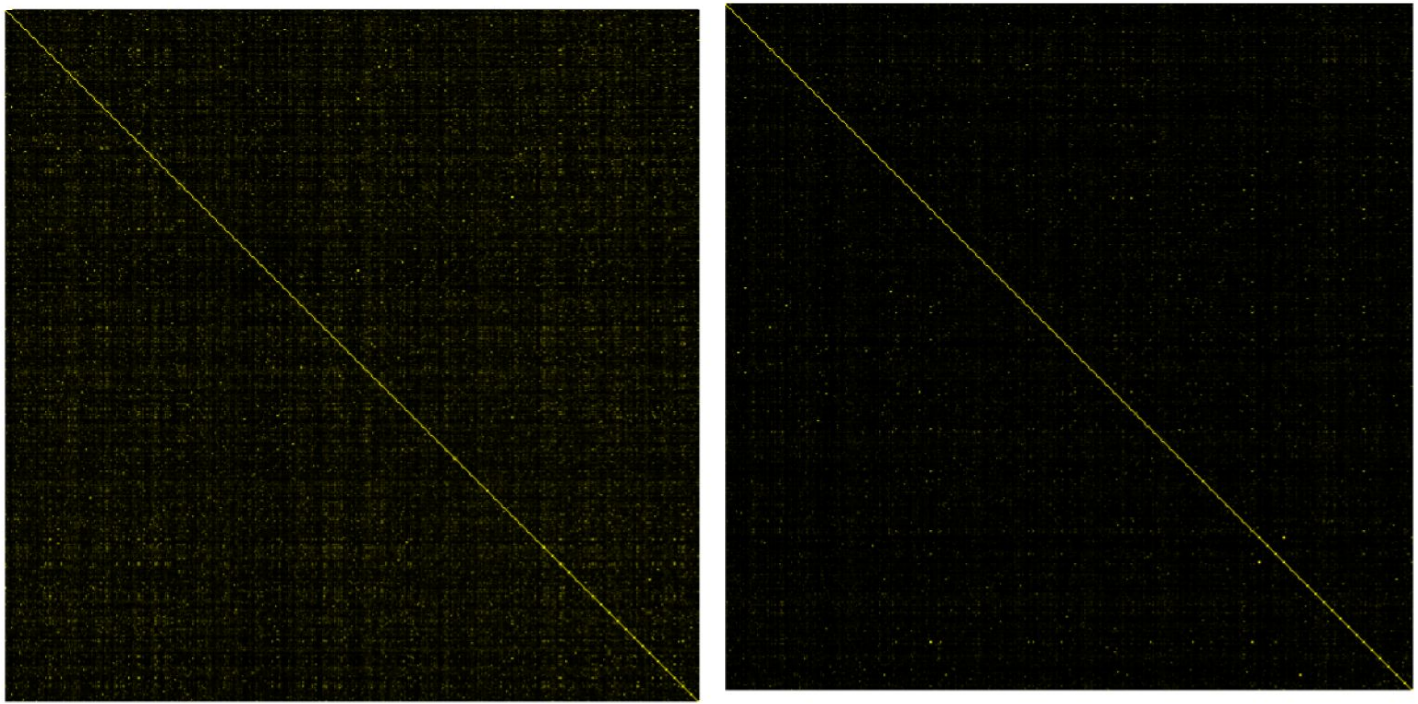


(a) English



(b) Chinese

Replication with larger documents for English



(a) English

(b) Chinese

Penn & Choma: Problems

- We attempted to replicate Penn & Choma's result using the Bible Corpus (Christodoulopoulos & Steedman, 2015), taking each chapter as a document
- This fails, and the reason seems to be because of the document sizes:
  - Brown corpus: 2000 words per document, i.e. about 4000 trigrams
  - Chinese news corpus (e.g. Chinese Gigaword) has about 450 characters per document
  - Bible:
    - Approx. 1100 trigram letters per document (chapter) for English
    - Approx. 780 characters per document for Chinese
    - Group 6 chapters into a "document" for English: about 6600 trigram letters per document
- A priori it seems unlikely that a measure just based on the distribution of symbols is going to be informative:
  - One needs to be able to relate them to pronunciation

Three proposals

- Simple lexical measure:
  - Count in a dictionary, or corpus, how many different spellings a given pronunciation has.
- Entropic measure
  - A logographic written symbol holds more information than a phonographic symbol.
  - Thus the conditional information (in the Shannon sense) should be lower
  - In other words the *conditional entropy* of a logographic system should be lower vis-a-vis the pronunciation than in a phonographic system
- Attention-based measure
  - In a neural attention-based sequence-to-sequence model *trained to spell words/morphemes from their pronunciations*, how much does the model need to *attend* to information in the context of the word?

Simple lexical measure

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

- *D* is a dictionary
- *s(p)* is the set of spellings for pronunciation *p*
- *C* is the corpus
- *c(p)* is the total count of each *p*

Entropic measure

This should be lower in a logographic system than in a phonographic system

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

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

$P_W(w)$ is the probability of (written) w given a written bigram model.

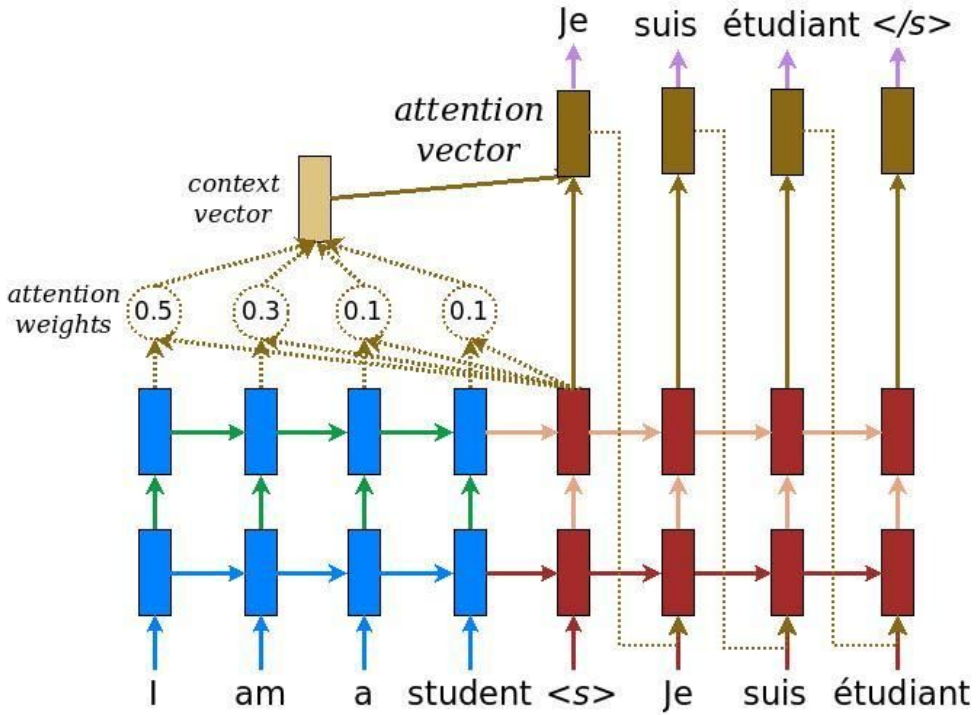
$P_P(p)$ is the probability of (pronounced) p given a pronunciation bigram model.

$H(X)$ is the entropy of variable X

$I(X, Y)$ is the mutual information between variables X and Y

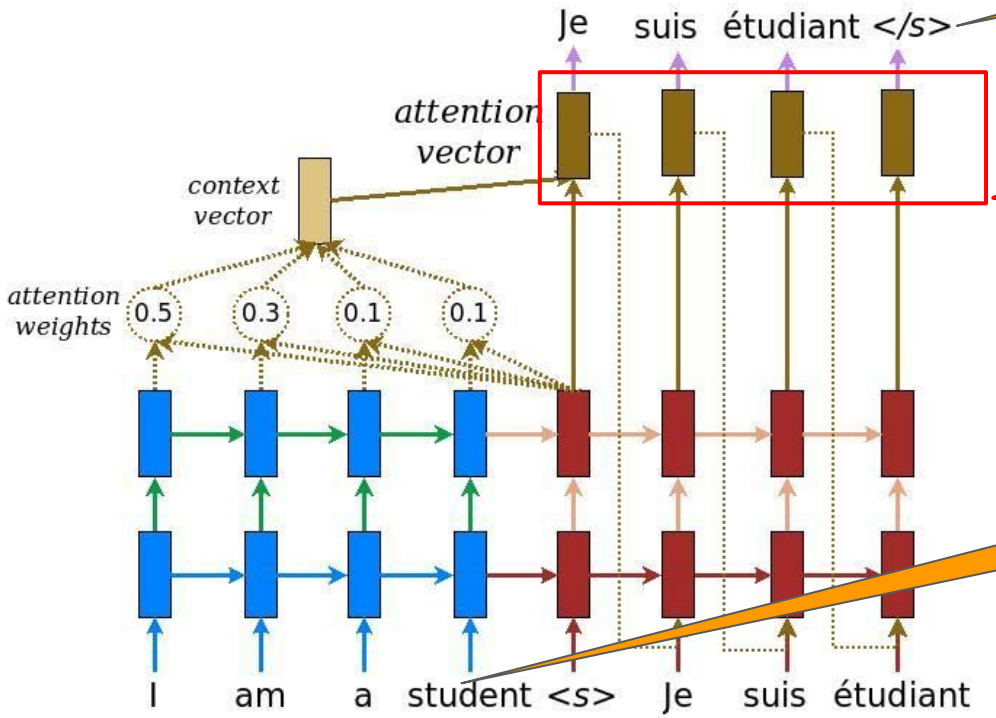
This should be lower in a logographic system than in a phonographic system

Attention based model



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

Attention based model



In our case: output is graphemes

Combine attention vectors into attention *matrix*

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

Finnish example

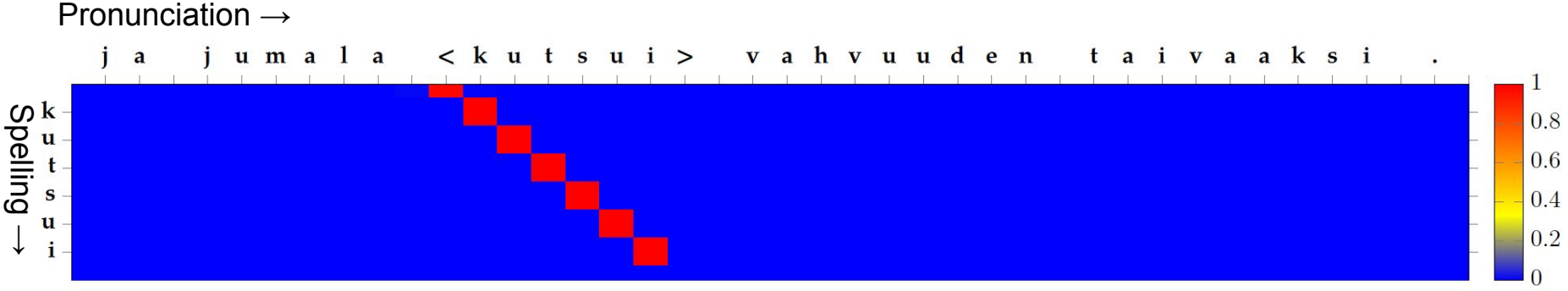


Figure 7
Attention matrix involved in spelling the Finnish word *kutsui* 'called'. The input (phonetic) sequence for the sentence is shown across the top of the plot, and the spelling of the target word is shown on the vertical axis. Note that in the plot itself the <targ> ... </targ> tags are reduced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

Finnish example

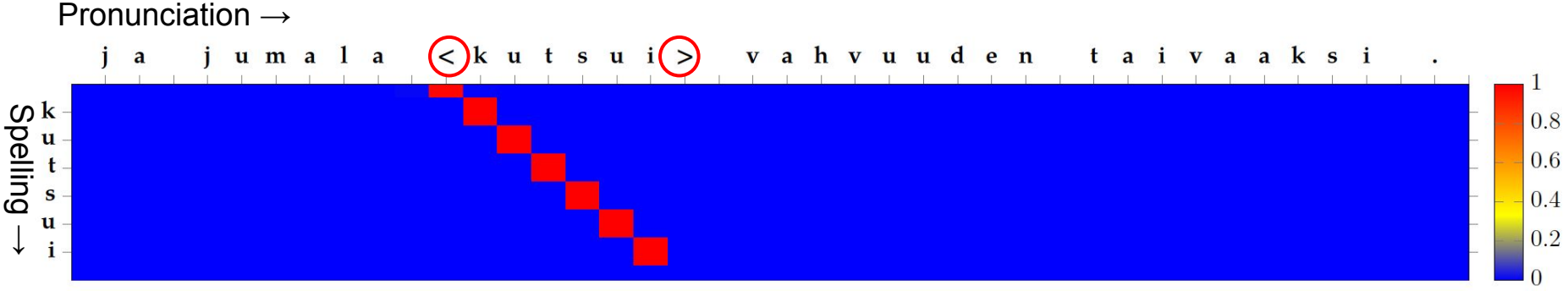


Figure 7
Attention matrix involved in spelling the Finnish word *kutsui* 'called'. The input (phonetic) sequence for the sentence is shown across the top of the plot, and the spelling of the target word is shown on the vertical axis. Note that in the plot itself the <targ> ... </targ> tags are reduced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

Chinese example

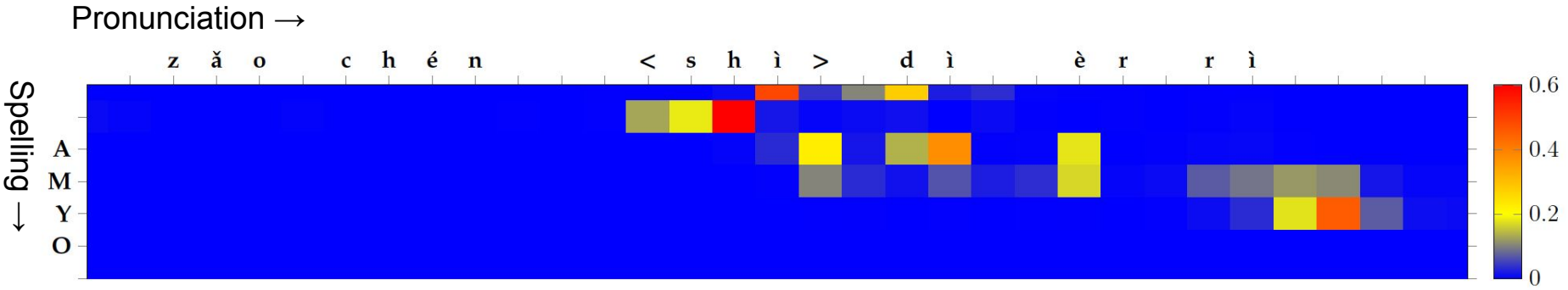


Figure 8
Attention matrix involved in spelling the Cangjie-encoded Chinese morpheme 是 (Cangjie AMYO) *shì* 'be'. (See Section 6 for details on encodings used for Chinese.) The input (phonetic) sequence for the sentence is shown across the top of the plot, and the spelling of the target word is shown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

Attention-based measure

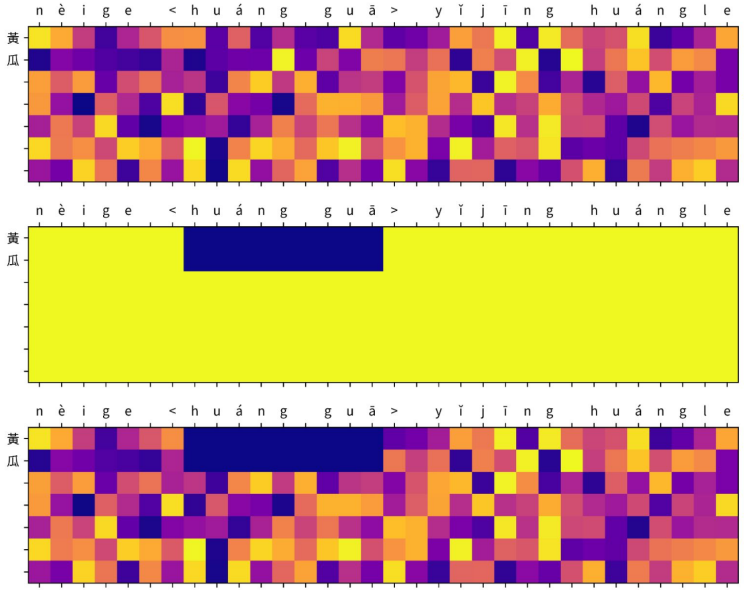


Figure 9
Illustration of the attention-based spread measure. Top: A random attention matrix. Middle: The zero mask for the target word. Bottom: The Hadamard product of the mask with the attention matrix.

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

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*

Data: Bible corpus

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

Jamo (individual Hangeul 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 | 。/.

Data

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

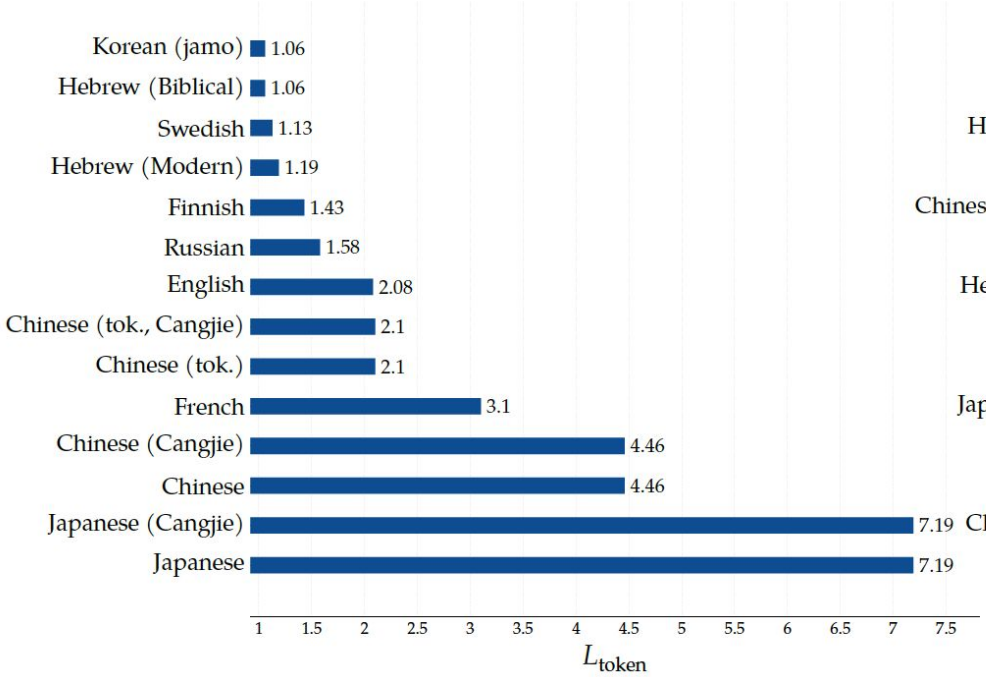
Results

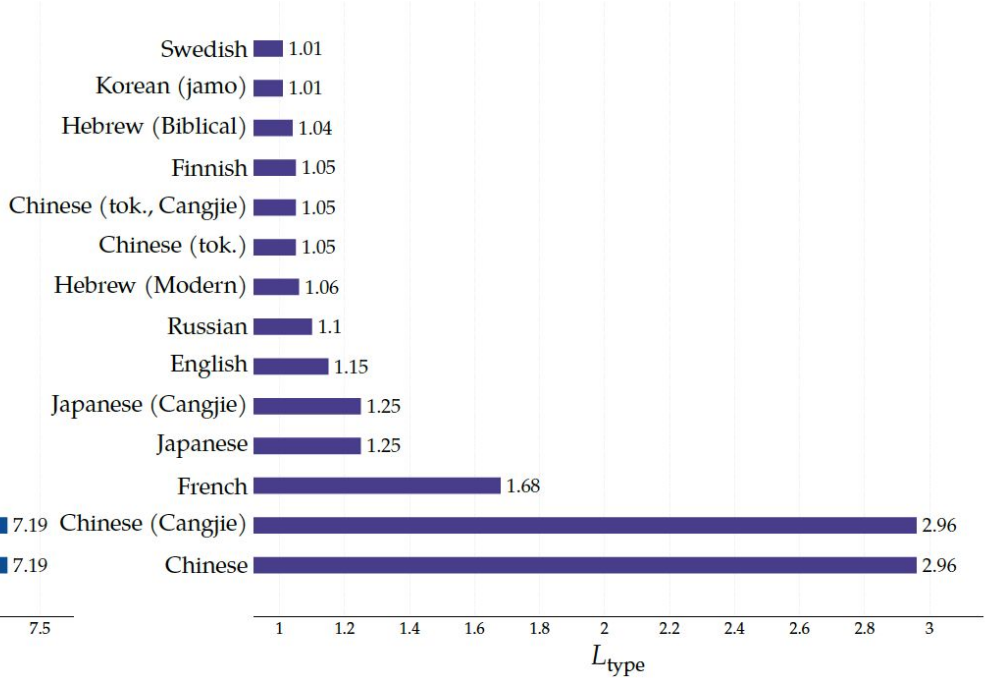
Language	Neural			Lexical		Entropic	
	S_{token}	S_{type}	Accuracy	L_{token}	L_{type}	E_{token}	E_{type}
Chinese	1.00	1.00	0.85	4.46	2.96	-0.12	7.86
Chinese (Cangjie)	0.74	0.71	0.87	4.45	2.96	-0.12	7.85
Chinese (tokenized)	0.55	0.37	0.89	2.10	1.05	-0.02	9.43
Chinese (tokenized, Cangjie)	0.51	0.32	0.78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

Results

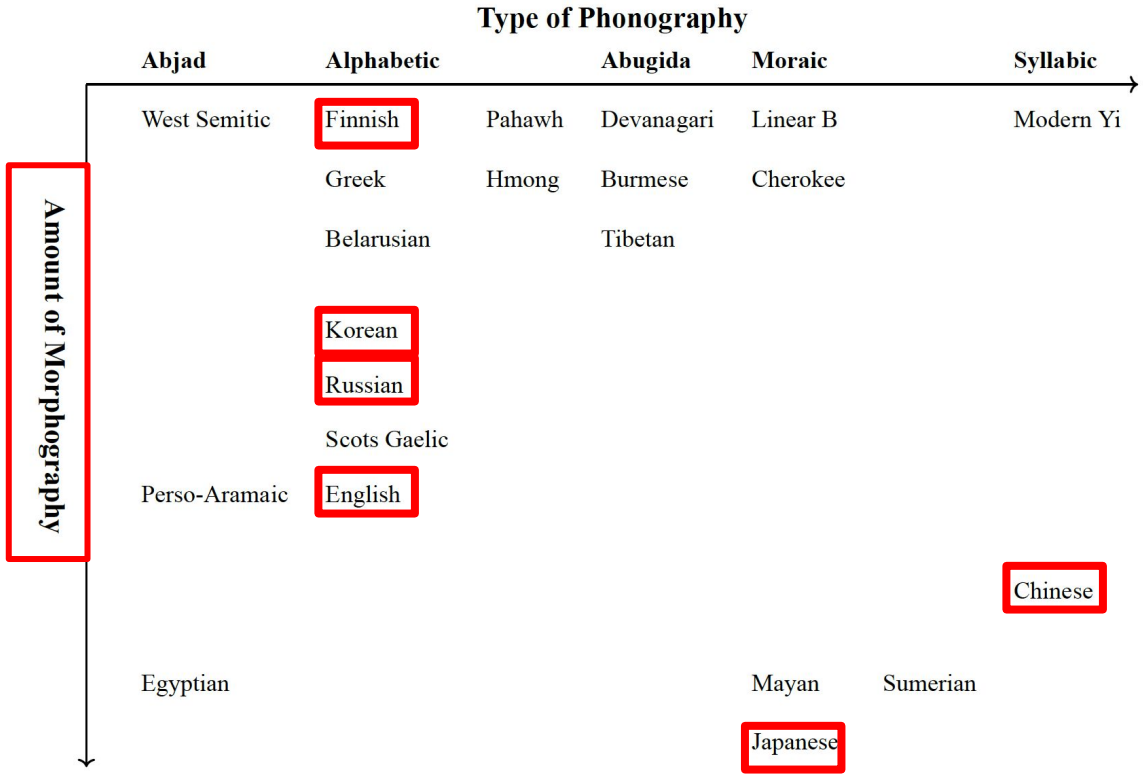
Language	Neural			Lexical		Entropic	
	S_{token}	S_{type}	Accuracy	L_{token}	L_{type}	E_{token}	E_{type}
Chinese	1.00	1.00	0.85	4.46	2.96	-0.12	7.86
Chinese (Cangjie)	0.74	0.71	0.87	4.45	2.96	-0.12	7.85
Chinese (tokenized)	0.55	0.37	0.89	2.10	1.05	-0.02	9.43
Chinese (tokenized, Cangjie)	0.51	0.32	0.78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

Results: Lexical measures





Background: taxonomies of writing systems



Rogers (2005) (< Sproat 2000)

Finnish

Korean

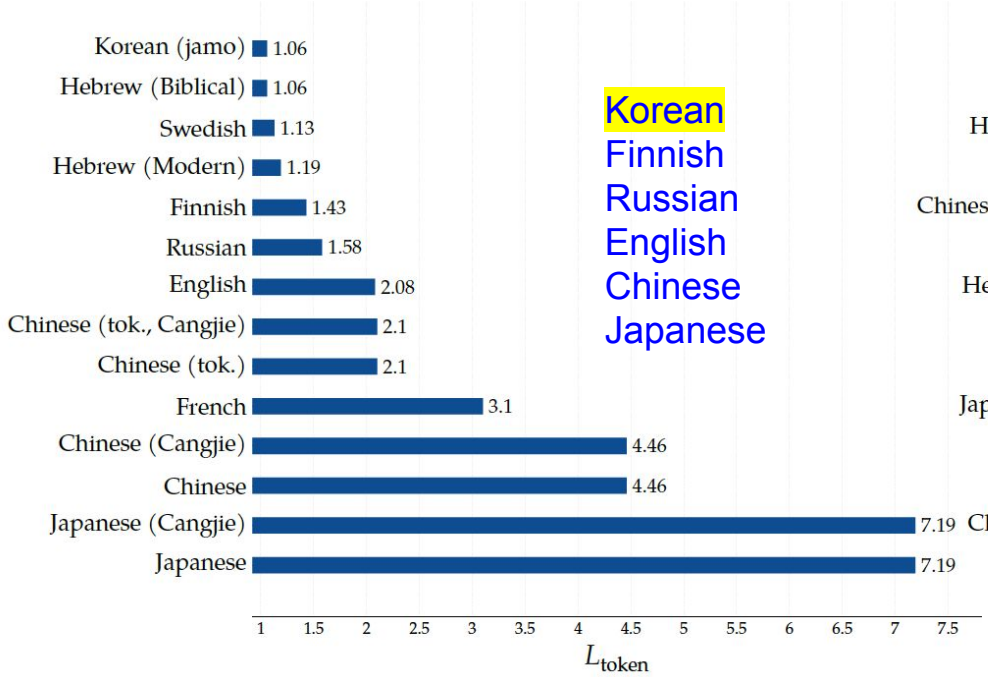
Russian

English

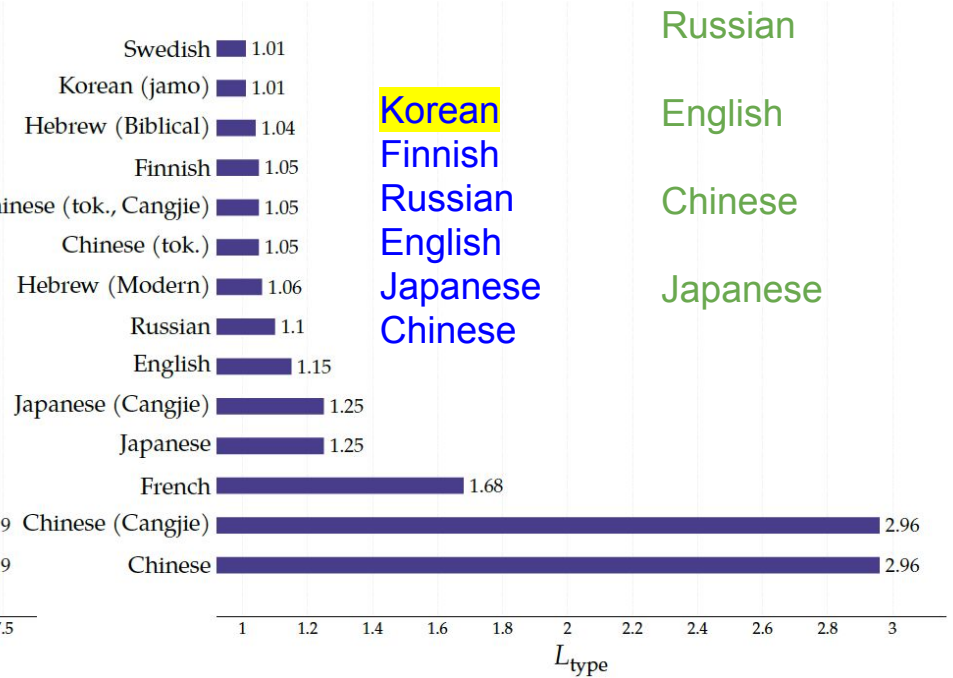
Chinese

Japanese

Results: Lexical measures



Korean
Finnish
Russian
English
Chinese
Japanese



Korean
Finnish
Russian
English
Japanese
Chinese

Finnish
Korean
Russian
English
Chinese
Japanese

Results: Entropic measures

Finnish

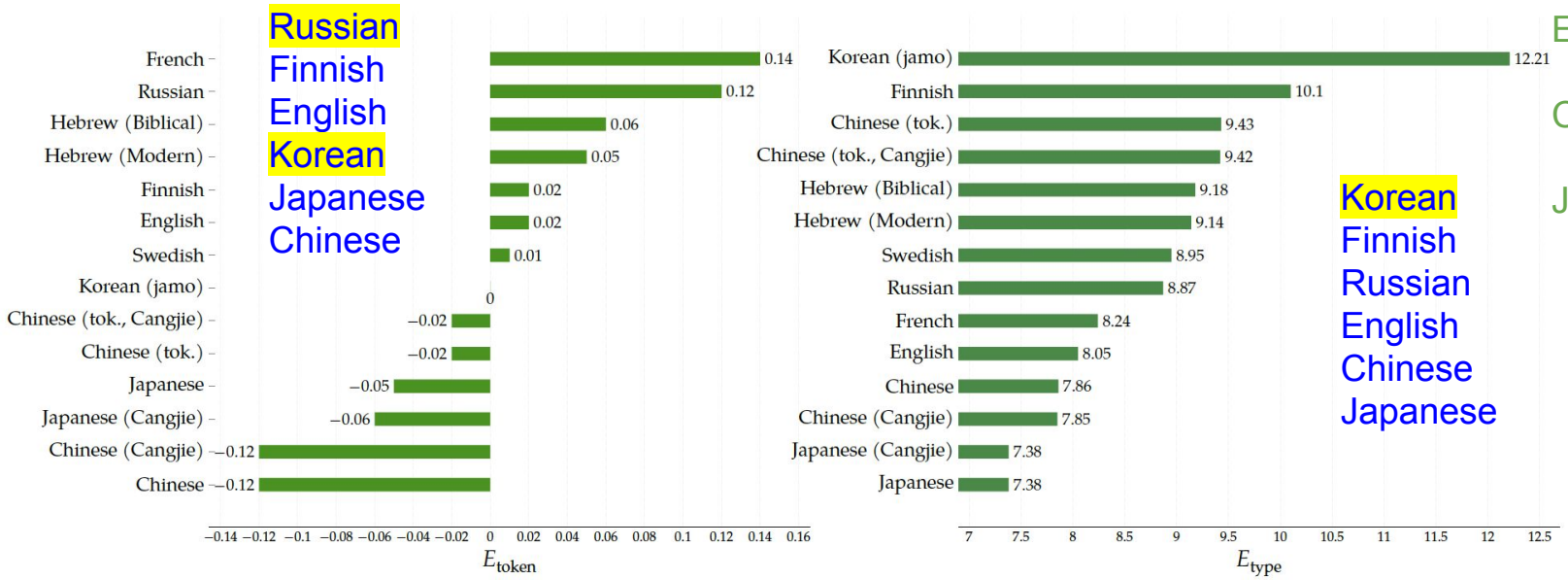
Korean

Russian

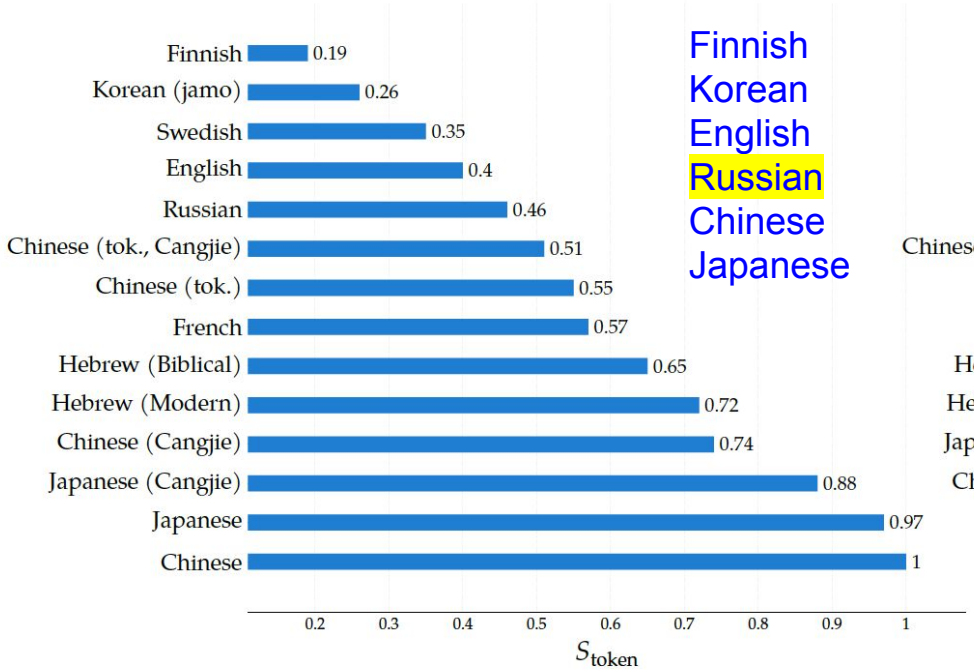
English

Chinese

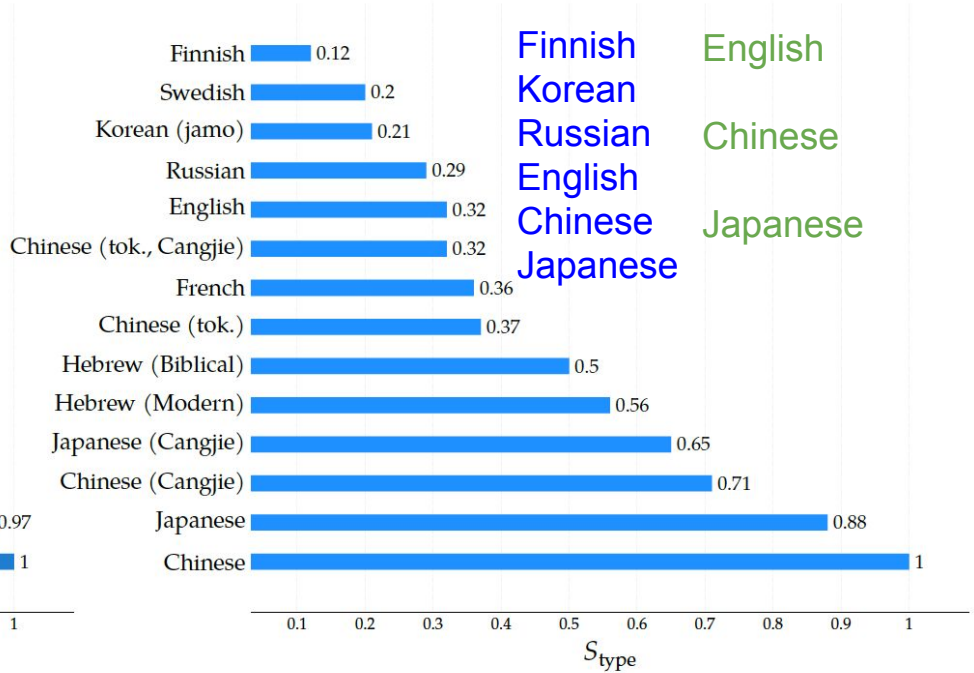
Japanese



Results: Attention-based measures



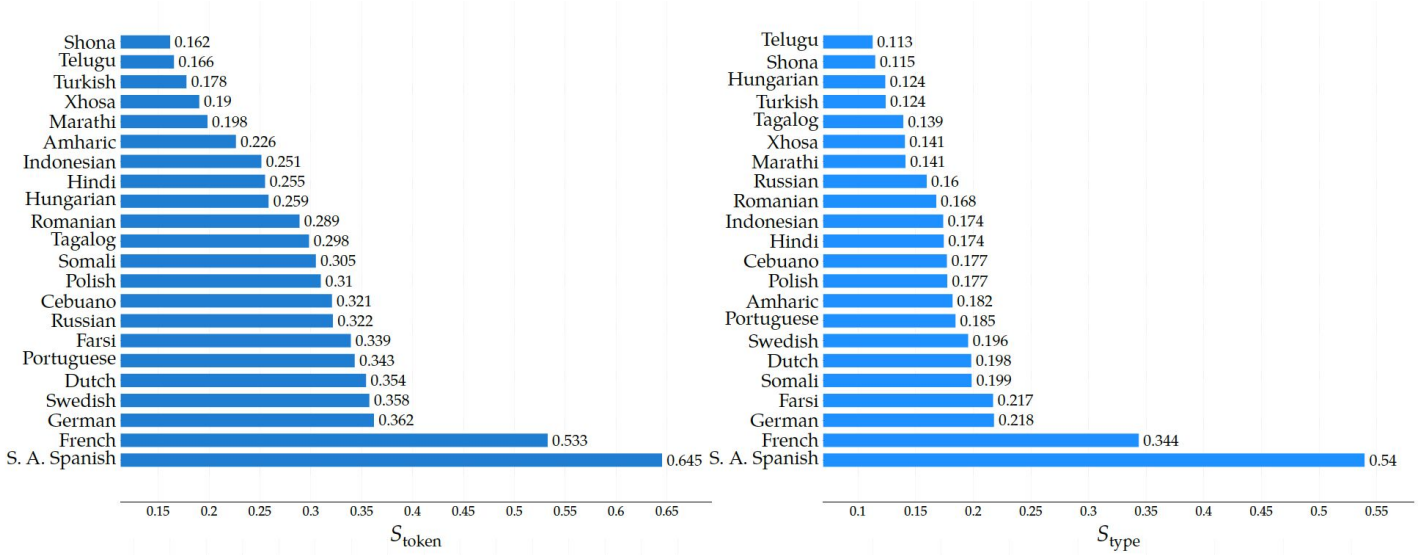
Finnish
Korean
Russian
Finnish
Korean
English
Russian
Chinese
Japanese



Finnish
Korean
Russian
English
Chinese
Japanese

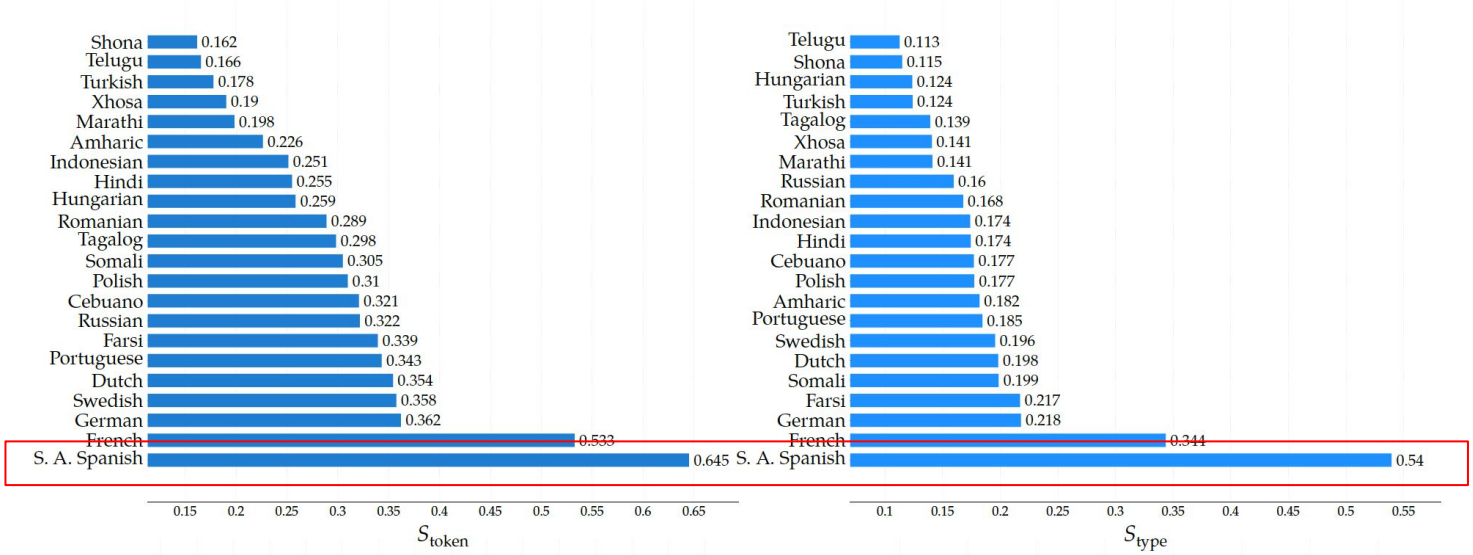
Additional experiments

- High(er) quality Japanese data: generally lower logography measure largely because of bigger tokens
- Epitran pronunciations



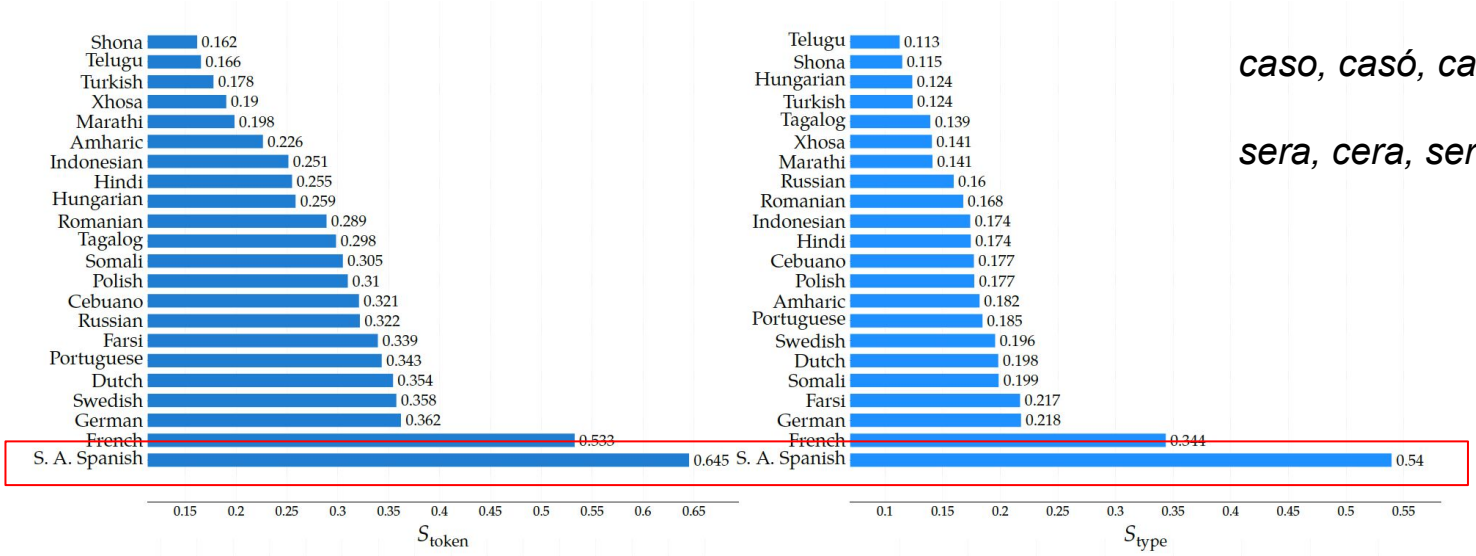
Additional experiments

- High(er) quality Japanese data: generally lower logography measure largely because of bigger tokens
- Epitran pronunciations



Additional experiments

- High(er) quality Japanese data: generally lower logography measure largely because of bigger tokens
- Epitran pronunciations



caso, casó, cazó

sera, cera, será

Critiques

- We have simply redefined the notion of “logography”
  - Hard to argue this since the notion has never really been defined before
- We’ve missed the point since in many (esp. ancient) logographic systems there are components of writing that clearly represent the meaning, not the pronunciation:
  - 琵琶 *pípá* “Chinese lute” vs. 枇杷 *pípá* “loquat”
  - 木 “tree” vs. 王王 “musical instrument” combined with 比巴 *biba*
  - Removal of one of these would render the example “non-logographic” by our measures
  - But such components are not critical to the notion of logography: cf. Sampson’s (1985) claim that English is at least partly logographic
  - One has to consider the behavior of the whole system
- All measures are sensitive to the data used

Images: Wikipedia.

Conclusions

- Attention-based measure seems to give intuitively satisfying results for the **distinct homophones** notion of logography:
  - How logographic a system is depends upon how much the writer must attend to the context to determine how to spell a word.
  - Other measures, in particular our entropic measures, also relate to that, but seem ultimately less satisfying.
- How logographic a system is depends upon the target of the spelling.
  - In the Chinese Bible, *dì* could be 6 different characters; *tiāndì* is only 天地 “heaven and earth”.
- By proposing a specific computational measure, we come to a better understanding of what “logography” means.
- Written paper is currently under review: please contact the authors if interested.