

A Computational Model of the Discovery of Writing

Richard Sproat
Google

CLASP
University of Gothenburg

May 14, 2018

Publication details

- Sproat, R. 2017. “A computational model of the discovery of writing”, *Written Language & Literacy*, 20:2, 194-226.
- Software released on Github


The problem

- There have been only four (more or less unequivocal) cases of the independent discovery of writing:
 - Mesopotamia
 - Egypt
 - China
 - Mesoamerica
- Why only there?

Some properties of all 4 systems

- All had methods for encoding *phonology*
 - They were not pure *logographic* or *semasiographic* systems
- Further, all had *mixed* systems employing some *semasiographic* and some *phonographic* symbols

鯉 *lī* 'carp' < 魚 FISH + 里 *lī*

ba  BALAM
'jaguar'
ma

r  w t   | *ršwt* 'joy'
š 


*urim*₅ + CITY = "Ur"₄

The problem

- There have been only four (more or less unequivocal) cases of the independent discovery of writing:
 - Mesopotamia
 - Egypt
 - China
 - Mesoamerica
- *With only four data points, it is hard to make any generalizations about the conditions that might favor (or disfavor) the discovery.*

Computational simulation

- Computational simulation has been used in a number of areas of linguistics to model phenomena that are hard to test in the laboratory:
 - Spread of linguistic features in social networks (e.g. Steels, 2012)
 - Historical change (e.g. Niyogi, 2006)
- *The present research seeks to model the emergence of writing from non-linguistic symbol systems*

Phenomena of interest

- The non-linguistic symbol systems in use in the culture, and the existence of combinatorial systems where symbols occur in “texts”.
 - Related issue: what *kinds* of non-linguistic systems would likely evolve into writing?

Combinatorial non-linguistic systems



Phenomena of interest

- The non-linguistic symbol systems in use in the culture, and the existence of combinatorial systems where symbols occur in “texts”.
- Linguistic properties favoring the discovery of writing, and favoring a particular kind of writing system over another (e.g. a consonantary versus a syllabary).
- Economic or other factors that would encourage the development of better means of record keeping.
- The development of lightweight materials encouraging the wider use of writing (Farmer et al. 2002).

Phenomena of interest

- The non-linguistic symbol systems in use in the culture, and the existence of combinatorial systems where symbols occur in “texts”.
- Linguistic properties favoring the discovery of writing, and favoring a particular kind of writing system over another (e.g. a consonantary versus a syllabary).
- Economic or other factors that would encourage the development of better means of record keeping.
- The development of lightweight materials encouraging the wider use of writing (Farmer et al. 2002).

What we will cover

- Phonological factors favoring the development of writing
- Two different pathways to “grammatogenesis”:
 - Symbols representing ideas versus symbols representing morphemes
- Was writing “invented”, as opposed to developing over time from non-linguistic symbol systems?

“Full”* writing and how to get there

In all early writing systems, a crucial insight was the realization that a symbol that had been used to represent a morpheme could be also be used to represent a different morpheme that sounds similar.

All full writing systems encode phonological information.

* “Full” meaning that one can write down pretty much anything one can say

The Model

Details

- About 1500 lines of Python code
- Grammars using *Thrax* finite-state grammar compiler (<http://www.openfst.org/twiki/bin/view/GRM/Thrax>)
- *Pynini* (<http://www.openfst.org/twiki/bin/view/GRM/Pynini>)
- Released on GitHub: https://github.com/rwsproat/writing_evolution

Main parameters for the program

<code>niter</code>	number of iterations
<code>ablaut</code>	whether to apply ablaut (Section 4)
<code>base_morph</code>	shape of base morph (Section 4)
<code>initialize_non primaries_with_symbol</code>	whether non-primary morphs associated with a concept should inherit the symbol (Section 5)
<code>freeze_semantics_at_iter</code>	freeze semantics at iteration n (Section 5)
<code>probability_to_seek_spelling</code>	probability of seeking a spelling (Section 6)

Phonology and phonotactics

Table 2. Basic phonemes of the systems. Notation: *uvstop* = unvoiced stop, *vstop* = voiced stop, *uvfric* = unvoiced fricative

	uvstop	vstop	uvfric	nasal
labial	P	b		m
dental	t	d	s	n
velar	k	g		N
	liquids		l, r	
	semivowels		w, y	
	vowels		a, e, i, o, u	

Parameters: phonotactics

Basic phonotactics for syllables:

$\sigma = (s? P? L? V (L | M)? P2)? | (s? M? V L? P2?)$

$c = (s | P | L)? V$

Table 3. Sample morphemes from each of the three morpheme structure conditions

monosyllable	σ	alp, alt, byun, byut, klilk, milk, newt, nort, prerik, rok
sesquisyllable	$c\sigma$	adiNk, agrot, astamp, aul, beuyp, duop, edu, gaek, gi, milk
disyllable	$\sigma\sigma$	awpblap, daykru, glutilt, gu, guNk, ilkuy, liak, lurtimp, prot, ratgla

Randomly generate $\approx 1K$ morphemes from these templates

Parameters: semantics

PERSON:♂, MAN:♂, WOMAN:♀, HOUSE:🏠, BRONZE:♁, GOLD:☉, SILVER:☽, SWORD:⚔, MEAT:🍖, SHEEP:🐏, OX:🐂, GOAT:🐐, FISH:🐟, TREE:🌳, BARLEY:🌾, WHEAT:🌾, WATER:💧, STONE:🪨, CLOTHING:👕, FIELD:🌾, TEMPLE:🏛️, GOD:👤, AXE:🔪, SCYTHE:🪓, DOG:🐕, LION:🦁, WOLF:🐺, DEMON:👹, SNAKE:🐍, TURTLE:🐢, FRUIT:🍏, HILL:🏔️, CAVE:🕒, TOWN:🏘️, ENCLOSURE:🏠, FLOWER:🌸, RAIN:☔, THUNDER:⚡, CLOUD:☁️, SUN:☀️, MOON:🌙, HEART:♥️, LUNG:🫁, LEG:🦵, ARM:🦶, FINGER:👉, HEAD:😊, TONGUE:👅, EYE:👁️, EAR:👂, NOSE:👃, GUTS:🗑️, PENIS:🍆, VAGINA:🍆, HAIR:👁️, SKIN:👁️, SHELL:🐚, BONE:🦴, BLOOD:🩸, LIVER:🍷, FARM:🏡, LOCUST:🦋, STICK:🪵, STAR:★, EARTH:🌍, ASS:🐴, DEATH:💀, BIRTH:👶, WOMB:👶, MILK:🥛, COAL:♁, SEED:🌱, LEAF:🍃, CHILD:👶, ANTELOPE:🐘, BEAR:🐻, BEE:🐝, MOUSE:🐭, DUNG:💩, PLOUGH:🌾, SPROUT:🌱, ICE:🧊, DAY:☀️, NIGHT:☾, WINTER:❄️, SUMMER:☀️, AUTUMN:🍂, SPRING:🌱, KING:👑, GOOSE:🐓, PRIEST:👤, ROAD:🛣️, CART:🛒, GRASS:🌿, FIRE:🔥, WIND:🌪️, NAIL:🔪, BREAST:👃, BOWL:🍲, CUP:🍵

Concepts, morphemes and symbols

- Generate about 1,000 morphs per run
- Randomly associate 1-3 morphs w/ a concept.
 - If `initialize_non primaries_with_symbol` is true, all morphs inherit the symbol associated with the concept
 - Otherwise only the “primary” morph does
 - HEART → *klik*, *bort*, with *klik* being “primary”.
 - ♥ can be used to write just *klik* if `initialize_non primaries_with_symbol` is true
 - or both *klik* and *bort* if it is false.
- Randomly associate remaining morphs to combinations of concepts, so that *mul* might map to AUTUMN, DAY

Parameters: random concept combos

CUP,ROAD,DAY

FARM,BLOOD,GOAT

EARTH,BONE,COAL

NIGHT,BEAR

DUNG,CHILD,DAY

Some of these combinations seem weird, but ...

Cf. Japanese *kokuji* 孀 /kaka/ “wife” = 女 “female” + 鼻 “nose”

Finding spellings

- In the initial setting, the system is *semasiographic* or *logographic*: symbols represent *concepts* or *morphemes*
- The system tries to find spellings for other morphemes, according to the setting of `probability_to_seek_spelling`
- Searches the lexicon for morphemes that
 - a. Share a meaning (e.g. DEMON,STONE could match DEMON), or
 - b. Sound similar

Parameters: what's phonetically close?

- Normalized edit distance computation where:
 - All segments match themselves freely
 - Segments can match other segments according to their phonetic distance: /p/ matches /b/ better than /p/ matches /k/
- *Telescoping* is also allowed, where /bak/ could be represented with two signs /ba+/ak/

Experiment 1: Different phonological conditions

Expt 1: Simulating different phonological conditions

Table 3. Sample morphemes from each of the three morpheme structure conditions

monosyllable	alp, alt, byun, byut, klilk, milk, newt, nort, prerik, rok
sesquisyllable	adiNk, agrot, astamp, aul, beuyup, duop, edu, gaek, gi, milk
disyllable	awpblap, daykru, glutilt, gu, guNk, ilkuy, liak, lurtimp, prot, ratgla

Ablaut. Optionally applied in disyllable condition

a→o

e→o

l→u

o→u

u→∅

Four conditions:

→ Monosyllable

→ Disyllable

→ Sesquisyllable

→ Disyllable with ablaut

Parameters

```
initialize_non_primaries_with_symbol=False  
probability_to_seek_spelling=0.5
```

Each condition was run 5 times, for 10 epochs each

Parameters: simulation



























1. *Assign simplex concepts and symbols to morphemes*
2. *Randomly assign complex concepts to other morphemes*
3. Try to assign spellings to morphemes that do not have one
 - a. Symbols associated with shared semantic components, *and/or*
 - b. Symbols associated with similar sound \Leftarrow “*eureka*” moment
 - i. Extend the phonetic coverage with *telescoping*:
 1. $ba + ad \rightarrow bad$
4. Result will be a mix of semantic and phonetic components

More on the “eureka” moment: phonological recoding of logographic signs

- Among modern writing systems, *kanji* hold the best claim to be *logographic*.
- Yet there is substantial evidence for “phonological recoding” of *kanji*.
 - E.g., Horodeck (1987) noted that many writing errors in Japanese involve substituting a character with the wrong meaning, but the right sound. Most of his examples involved *on* (Sino-Japanese) readings, though 7% involved *kun* (native) readings.
- *Conclusion: phonological recoding is more or less automatic among fluent readers.*

	<i>Ma.</i>	<i>Ja.</i>	<i>(kun)</i>	<i>Meaning</i>
里	lǐ	sato		⅓ mile, village
鯉	lǐ		koi	carp
裡	lǐ	ura, uchi		inside
理	lǐ	suzi, kotowari, osameru		reason
厘	lí	mise		1000th of a tael

Some sample generated spellings

	<i>riwk</i>	ASS	( ASS)
	<i>yint</i>	TEMPLE,SKIN	( TEMPLE +  <i>int</i>)
	<i>mol</i>	COAL	( <i>mo</i> +  <i>ol</i>) (< AUTUMN <i>mo</i> + NAIL <i>ol</i>)
	<i>mol</i>	SPROUT	( SPROUT + <i>mol</i>  
	<i>up</i>	SNAKE,SPROUT,DUNG	(   SPROUT)
	<i>up</i>	TEMPLE,FLOWER	( FLOWER +    <i>up</i>)
	<i>rulp</i>	COAL	( <i>ru</i> +    <i>up</i>)

Results

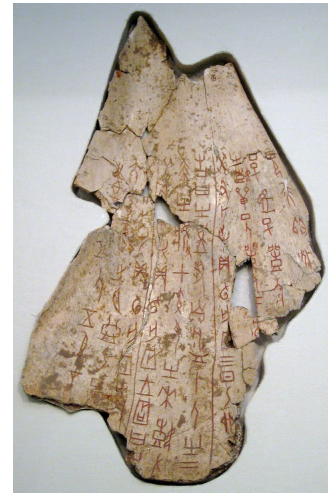
Table 4. Results from the first experiment. ‘Tot’ is the total mean proportion of morphemes with a spelling, across the various runs for the condition; Φ is the proportion of spellings that are purely phonetic; $S\Phi$ those that are semantic- phonetic; and S purely semantic spellings. The first row is the means, the second the standard deviations

monosyllable				sesquisyllable				disyllable				disyllable + ablaut			
Tot	Φ	$S\Phi$	S	Tot	Φ	$S\Phi$	S	Tot	Φ	$S\Phi$	S	Tot	Φ	$S\Phi$	S
0.81	0.23	0.32	0.45	0.35	0.12	0.20	0.67	0.34	0.12	0.18	0.70	0.45	0.17	0.25	0.58
0.03	0.02	0.03	0.02	0.02	0.01	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01

DeFrancis, 1984

Table 3 Structural Classification of Characters

Principle	Oracle Bones (Shang dynasty)	Xu Shen (2nd century)	Zheng Qiao (12th century)	Kang Xi (18th century)
Pictographic	227 (23%)	364 (4%)	608 (3%)	±1,500 (3%)
Simple indicative	20 (2%)	125 (1%)	107 (1%)	
Compound indicative	396 (41%)	1,167 (13%)	740 (3%)	
Semantic-phonetic	334 (34%)	7,697 (82%)	21,810 (93%)	47,141 (97%)
Total	977	9,353	23,265	48,641



Evolution over epochs

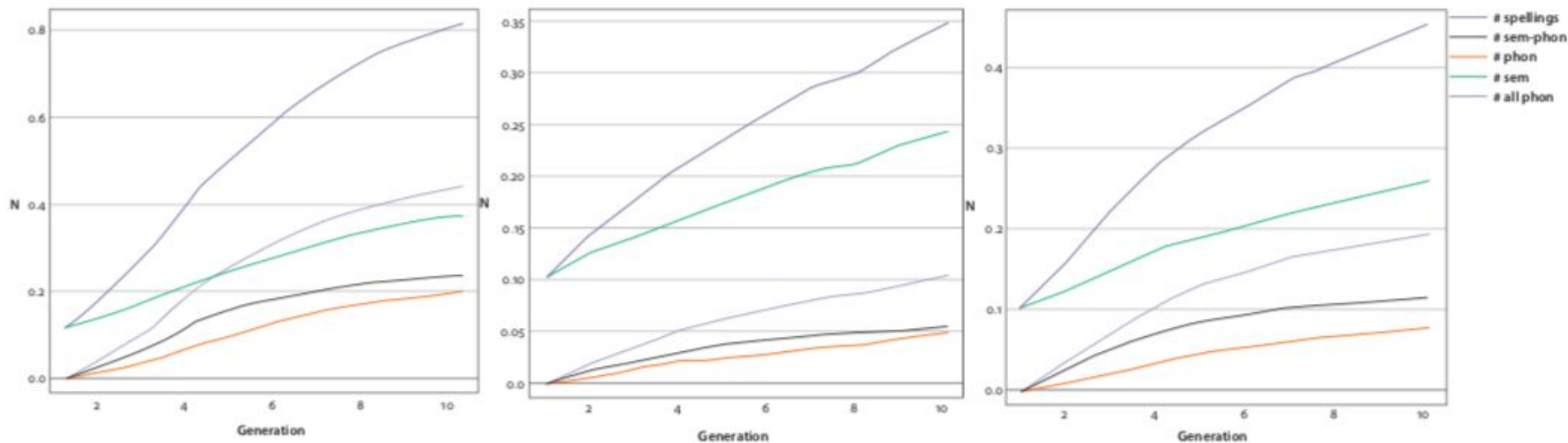


Figure 4. Results showing the evolution for one of the simulations for each of the **monosyllable**, **disyllable** and **disyllable with ablaut**. Shown are the growths in the proportion (N) of spellings (relative to the total number of morphemes), the total proportion of semantic spellings, the proportion of semantic-phonetic and pure phonetic spellings, and the total proportion of phonetic spellings.

Experiment 2:


**Symbols for concepts vs
Symbols for morphemes**

Expt 2: Symbols representing concepts vs morphemes








Table 5. Characters with the phonetic component 丘 and their Old Chinese pronunciation according to Baxter and Sagart (2014)

Char.	Phonetic component	Mand.	Middle Chinese	Old Chinese
丘	丘	qiū	khjuw	k ^{wh} ə
蚯	丘	qiū	khjuw	k ^{wh} ə
虛	丘	xū	khjo	q ^h a
岳	丘	yuè	ngæwk	ŋ ^ʳ rok

Table 6. Example of the Sumerian symbol A₂, from (ETCSL, 2006). The left column is the conventional transcription for the symbol, the center column the actual cuneiform symbol, and the right column its various phonetic uses

A₂ |  | a₂, ed, et, id, it, it, te₈

Some more Sumerian examples

Sign names	Signs	ETCSL values
A		a, dur ₅ , duru ₅
A.AN		am ₃ , em _x , šeg ₃
A.EDIN.LAL		ummud
A.HA.TAR.DU		girim ₃
A.IGI		er ₂ , še _x
A.KA		ugu ₂
A.KAL		illu

Two neurological routes to grammatogenesis

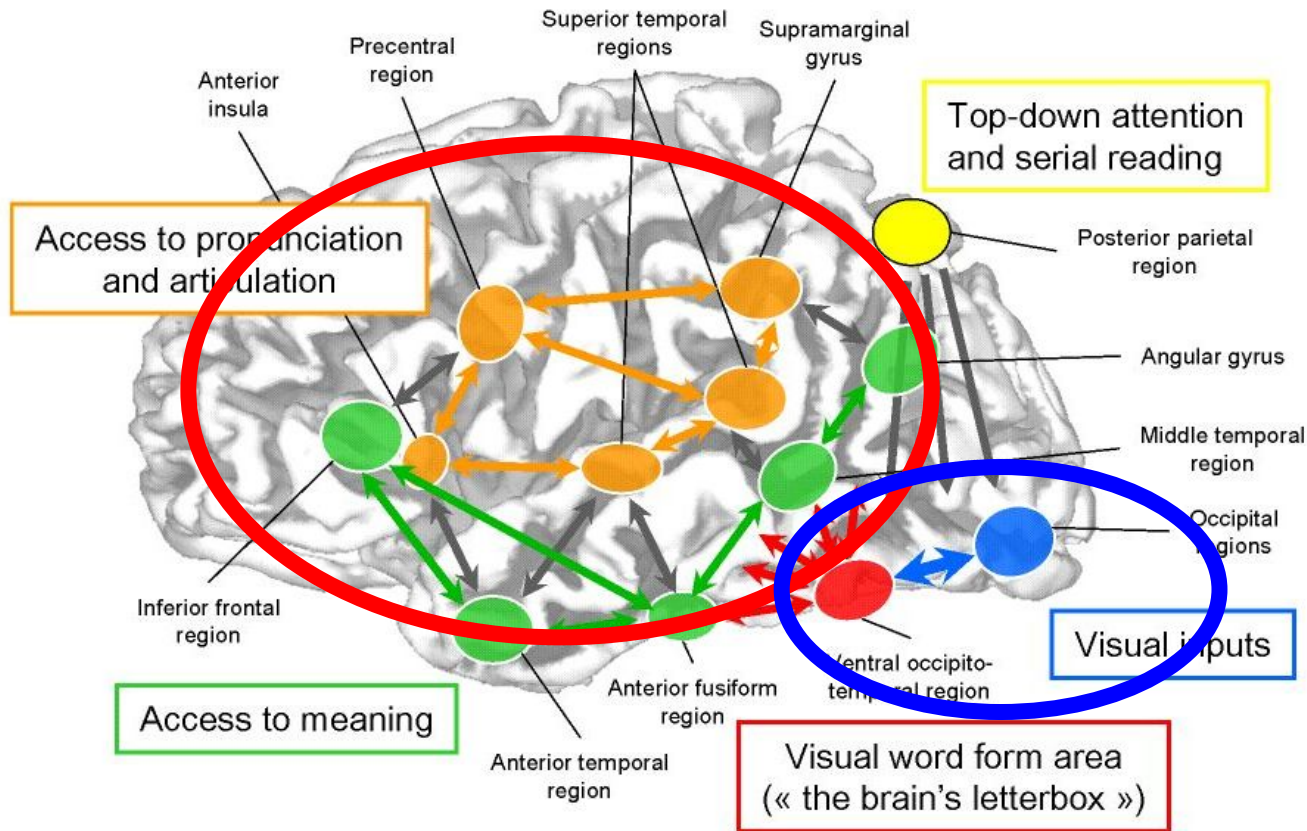
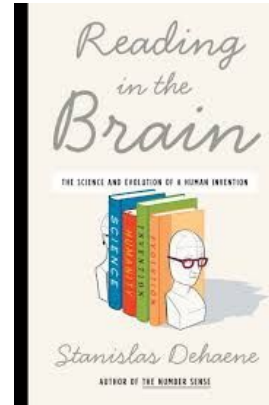


Image from:

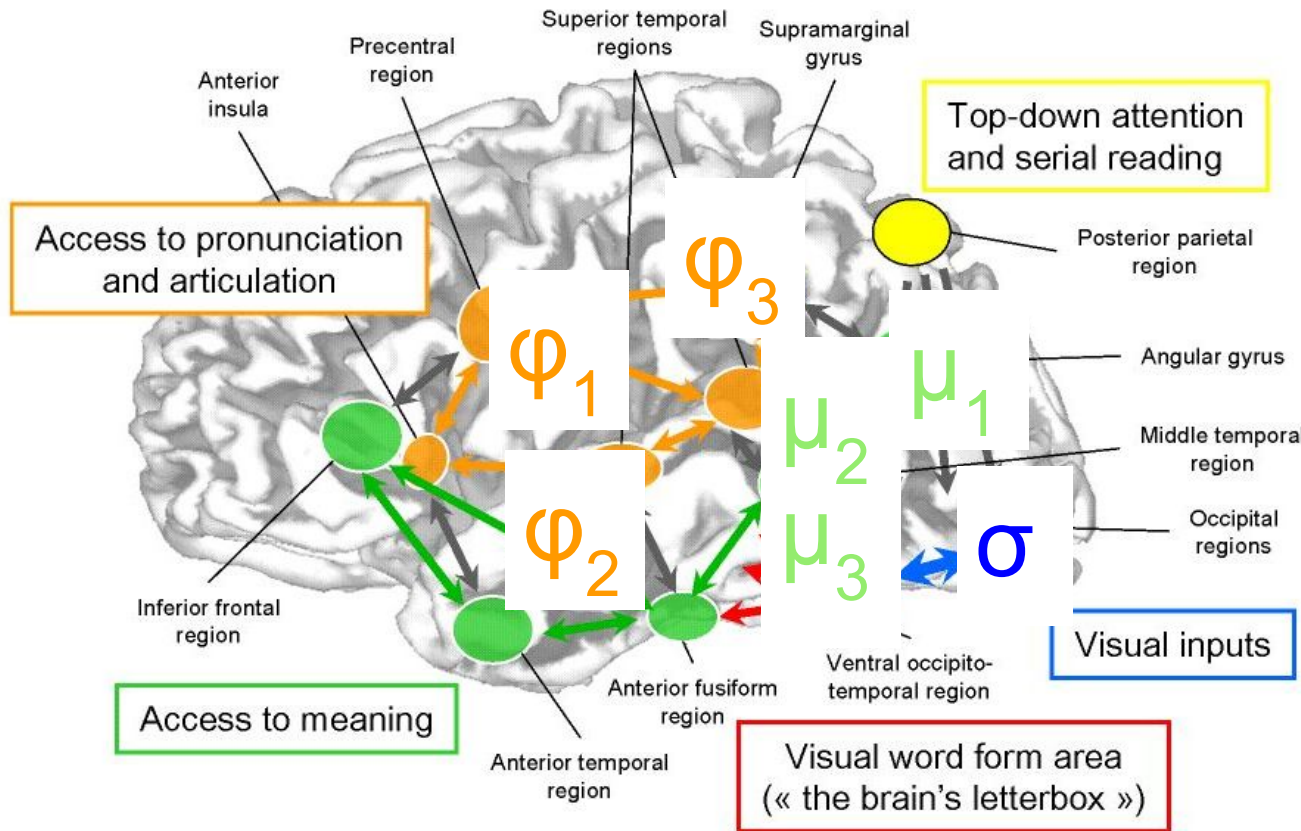
Dehaene, Stanislas. 2010. *Reading in the Brain*. New York, Penguin. Fig 2.2. p. 63



Language and speech

Visual processing

Two neurological routes to grammatogenesis

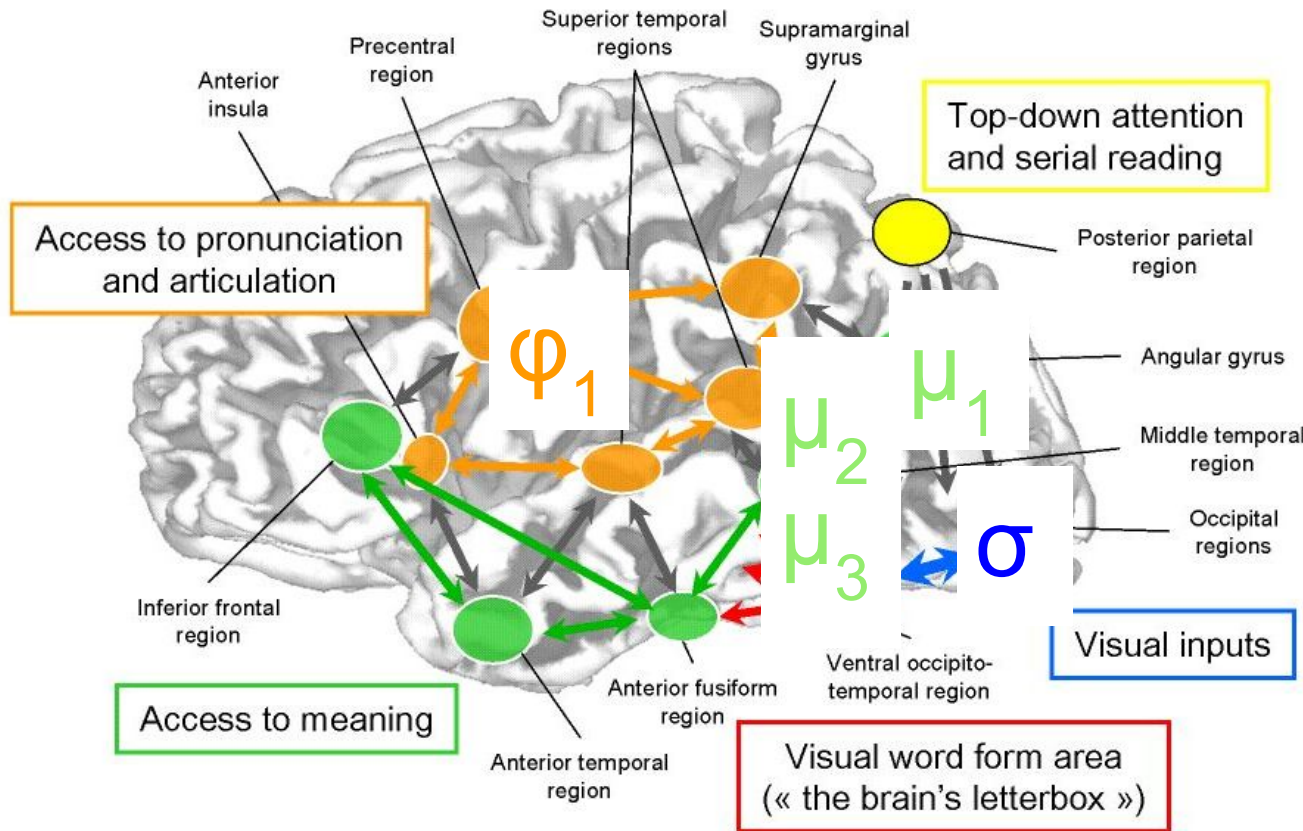


Symbol σ associated with concept and thence a set of related meanings/morphemes

μ_k

Via the morphemes becomes associated to set of pronunciations φ_k

Two neurological routes to grammatogenesis



Symbol σ associated with concept and thence a set of related meanings/morphemes

μ_k

One morpheme μ_1 becomes most strongly associated with the symbol.

Via this morpheme becomes associated to pronunciation ϕ_1

Two neurological routes to grammatogenesis


- The first effectively “fossilizes” the non-linguistic origin of the sign, preserving it through to multiple phonetic functions. **This is like Sumerian.**
- The second treats the sign as linguistic earlier by associating it to a particular morpheme and thence to a particular sound. **This is like Chinese (or Egyptian).**
 - *The second seems to reflect a more advanced stage: the inventors of the system realize that a sign can stand for a particular abstract linguistic unit.*

Phonetic uses for symbols

Table 5. Characters with the phonetic component 丘 and their Old Chinese pronunciation according to Baxter and Sagart (2014)

Char.	Phonetic component	Mand.	Middle Chinese	Old Chinese
丘	丘	qiū	khjuw	k ^{wh} ə
蚯	丘	qiū	khjuw	k ^{wh} ə
虛	丘	xū	khjo	q ^h a
岳	丘	yuè	ngæwk	ŋ ^ʳ rok

Table 6. Example of the Sumerian symbol A_2 , from (ETCSL, 2006). The left column is the conventional transcription for the symbol, the center column the actual cuneiform symbol, and the right column its various phonetic uses

A_2 |  | $a_2, ed, et, id, it, it, te_8$

Phonetic uses for symbols

1. For each distinct phonetic symbol k in the set V of phonetic symbols used in more than one morpheme:
 - a. For each pair of distinct phonetic values p_i, p_j of k , where Lev is the normalized Levenshtein distance:

$$subtot_k = \sum_{i,j|i < j} Lev(p_{i,k}, p_{j,k})$$

- b. For N_k = the number of distance computations performed for k

$$tot = \sum_k \frac{subtot_k}{N_k}$$

2. Return $divergence = \frac{tot}{|V|}$

(For a perfectly regular phonetic system $divergence = 0$)

Chinese vs. Sumerian

- Chinese: 1,102 phonetic symbols from Baxter & Sagart's (2014) reconstruction of Old Chinese
 - Phonetic divergence: **0.57**
- Sumerian: 212 symbols from the Electronic Text Corpus of Sumerian Literature (2006)
 - Phonetic divergence: **0.89**

One caveat: 'Old Chinese' is about 500 years later than the Oracle Bone Inscriptions

Simulation

- Toggle flag

`initialize_non primaries_with_symbol`

- We also need to freeze the semantic extension

`(freeze_semantics_at_iter, set to epoch 2)`

- Cf. the set of semantic “radicals” in Chinese which is about 200 and hasn’t changed much over 2000 years...
- vs. set of characters used as phonetic values in other characters, which is around 850 and has grown over time.

Results

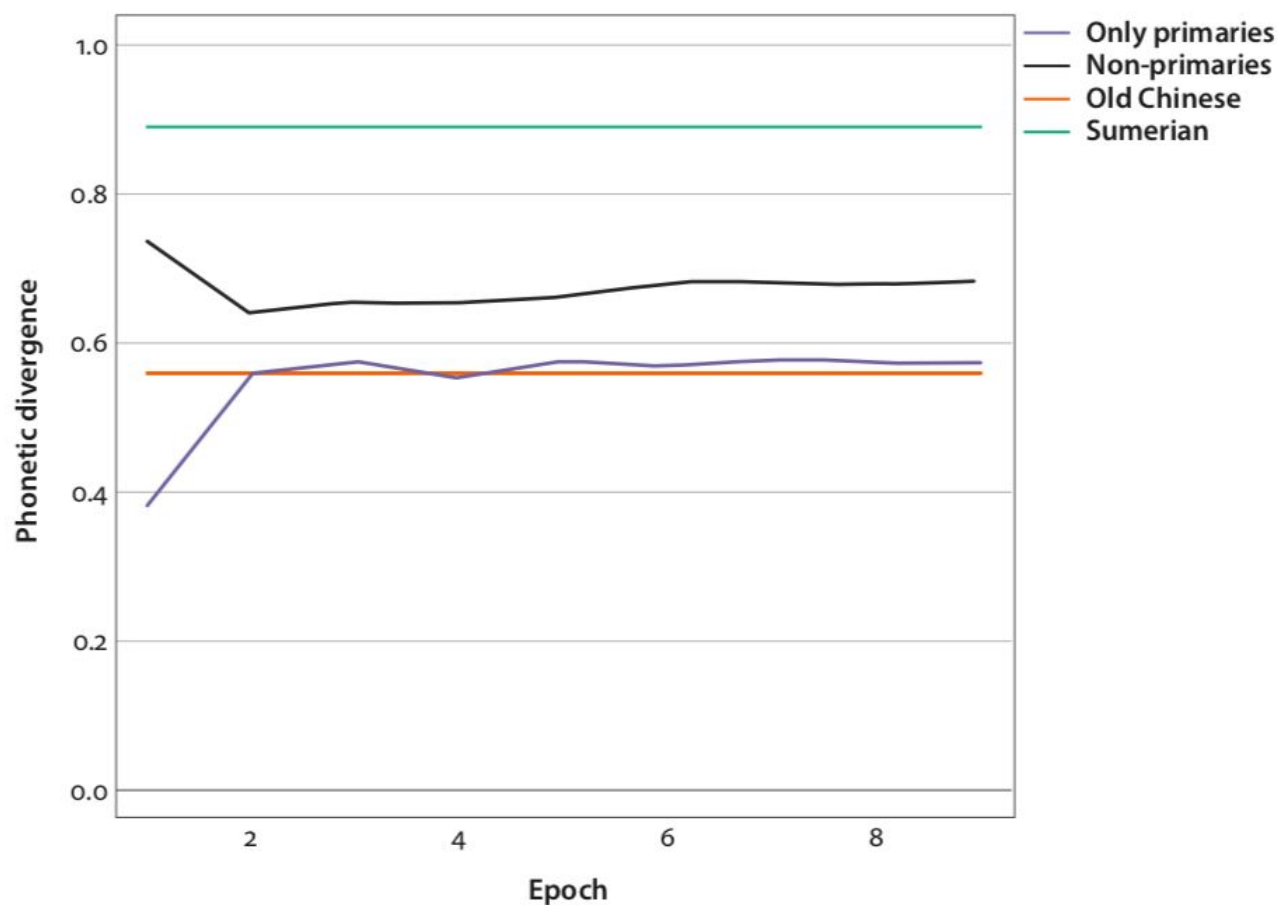


Figure 6. Evolution of phonetic divergence for phonetic equivalence classes in two monosyllabic systems, under the assumption that the further extension of semantic categories is “frozen” after epoch 2.

Summary of Experiment 2

- A model where one morpheme is picked as *the* denotation of the symbol fits Chinese better.
- A model where pronunciations spread from all morphemes associated with a concept fits Sumerian better.

Summary of experiment 2

- Does this reflect a difference in the evolution of the scripts?
- Sumerian developed from a raw “ideographic” system.
- But Chinese had symbols already associated with specific morphemes — *a later phase in the evolution of writing?* Either:
 - Phonetics were standardized from an earlier system...
 - or maybe Chinese got the idea of writing from elsewhere...
(see Boltz, 1994, 2004 for discussion, *though he does not support this conclusion*)

Experiment 3:

Was writing “invented”?

Expt 3: Was writing “invented”?

- The idea that writing evolved from non-linguistic symbol systems could be taken to be uncontroversial, but for the work of Glassner (2000):
 - Writing was *consciously* invented by its inventors and
 - “*Il ne peut y avoir, par définition ni pré- ni proto-écriture, ni écriture en gestation*”

JEAN-JACQUES GLASSNER

ÉCRIRE À SUMER

L'invention du cunéiforme



L'UNIVERS HISTORIQUE
SEUIL





Glassner's arguments

- Defects in the received theories of the origins of writing:
 - The “pictographic” theory, by which Glassner means narrative pictographic systems such as those used by Native Americans:



- The “Token” theory: Oppenheim (1959), Schmandt-Besserat (1996).
- Sumerian shows its “phonetic” character from the earliest times

Criticism of Glassner

- Glassner's theory has been critiqued by several scholars (Dalley, 2005; Robson, 2005; Englund, 2005)
- Glassner's argument about narrative pictographic systems is largely beside the point:
 - No evidence that such systems have *ever* been the basis of writing
 - On the other hand there is no question that symbols in various writing systems clearly did have pictographic origins:
 - Cf. Chinese 馬 'horse' <  龜 'turtle' < 
- Glassner's real issue comes down to the rapidity with which Sumerian writing showed *phonetic* properties

Glassner's point about phonetics

- A bit misleading since phonetic properties are usually taken as a defining characteristic of full-fledged writing.
 - I.e. a “pre-phonetic” symbol system would not be considered writing
- So the issue really comes down to the speed
 - The apparently sudden appearance of a full-fledged system could not be explained by natural processes of cultural evolution
- This is an error of the same kind as has plagued biological evolution and what Darwin called “organs of extreme perfection”: how did the vertebrate eye evolve?
 - Cf. Gould (1974, 104): “The dung-mimicking insect is well-protected, but can there be any edge in looking 5 percent like a turd?”

Simulations

- Vary `probability_to_seek_spelling`
- Simulates the amount of pressure to find ways to spell new words or morphemes

Results

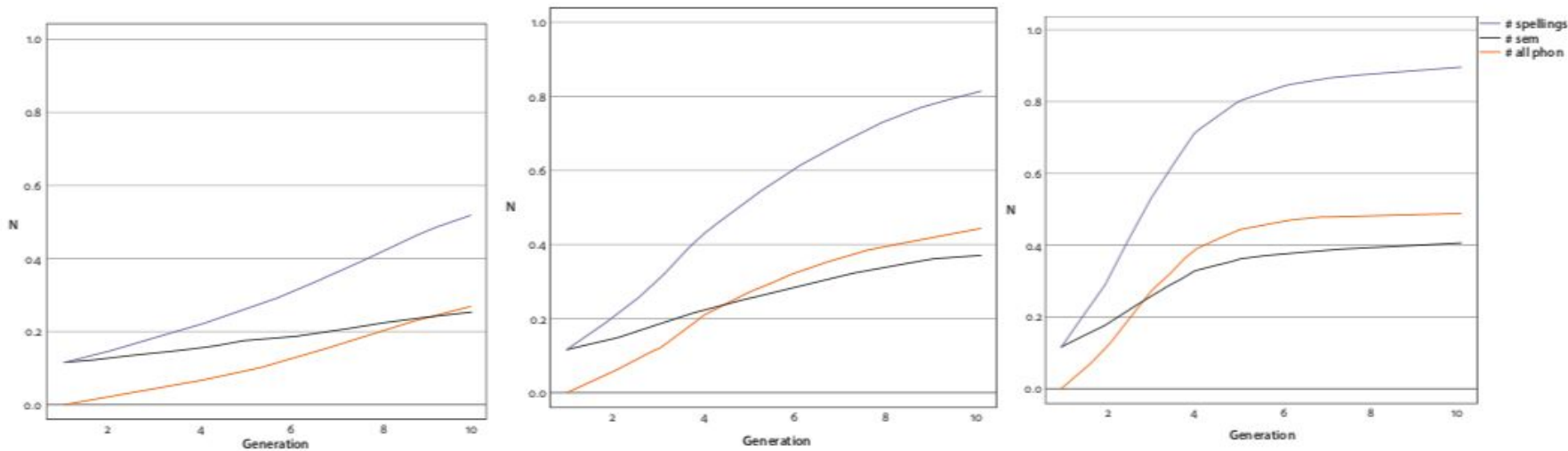
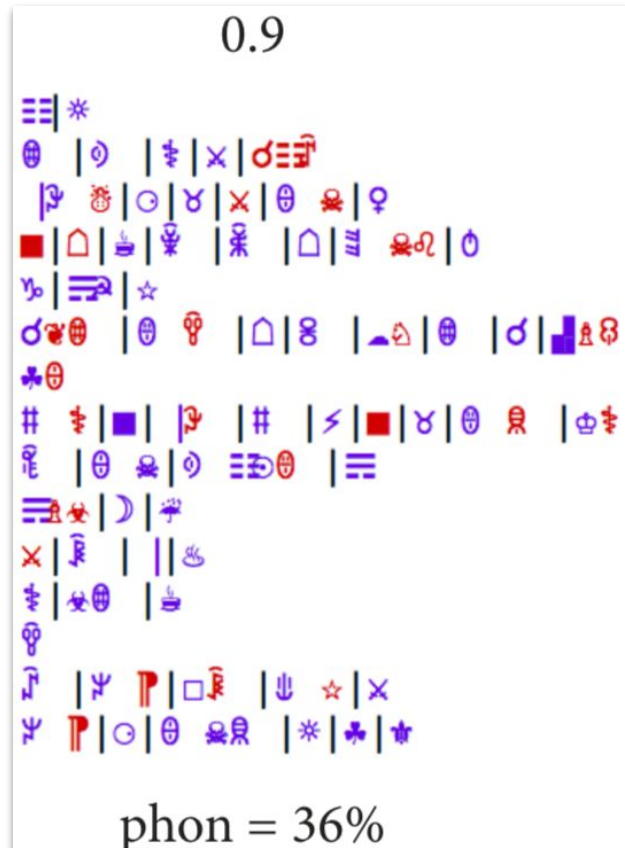
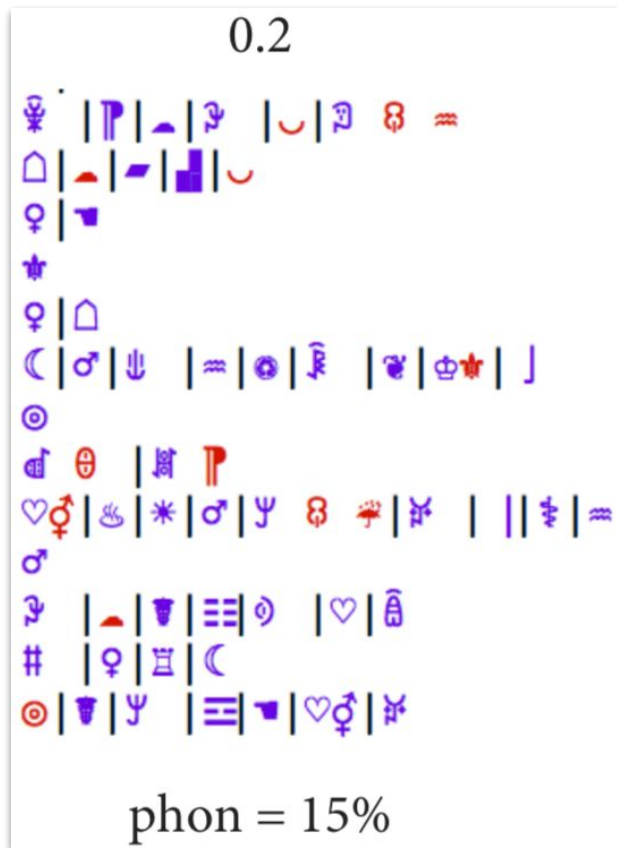


Figure 7. The evolution of writing in three monosyllabic languages with different probabilities of seeking a spelling for morphemes: 0.2, 0.5, 0.9. Again, the vertical axis N is the proportion of morphemes that have a spelling.

Sample “texts” at the second epoch



Summary

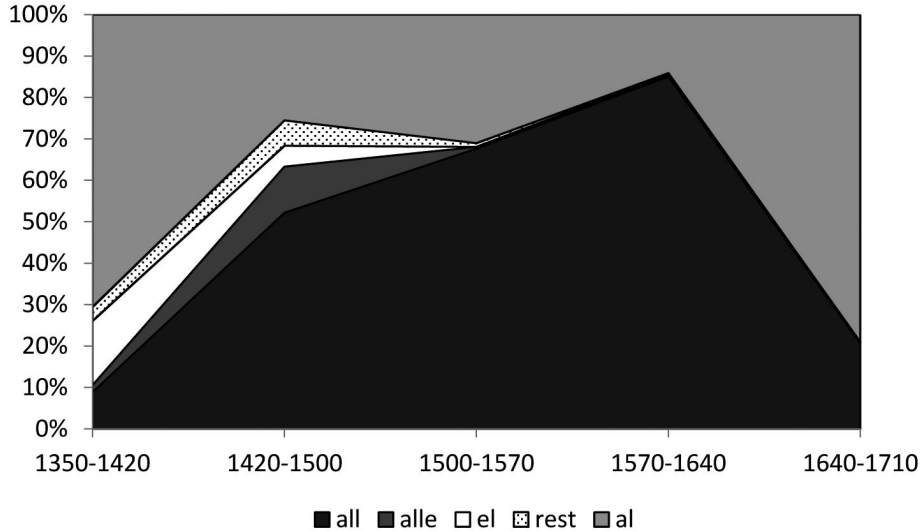
Summary

- Phonological form is important (Experiment 1):
 - Languages with largely monosyllabic morphemes have an edge (Steinthal, 1852; Daniels, 1992; Boltz, 2000; Buckley, 2008)
- Whether you start with a “mature” system makes a difference in how the system evolves (Experiment 2)
- Writing was not “invented” (Experiment 3)

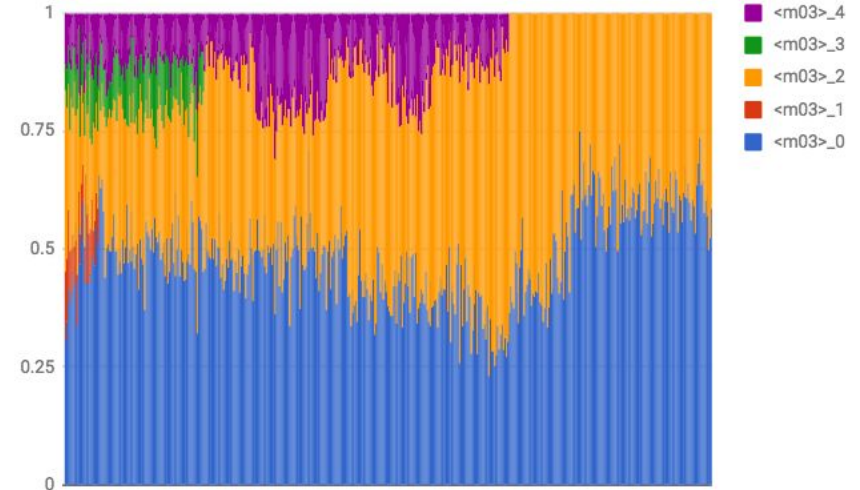
Further work

- Simulate a wider range of phonetic shapes
- More work on ablaut-like processes ... we need a better explanation for Egyptian
- Provide a more plausible model of lexical statistics
- Simulate writing of morphology:
 - so far we have simulated writing isolated morphemes
- Set of symbols is currently static:
 - new symbols are invented in real writing systems
- Models of standardization of spelling

Standardization



Proportions of spellings of *-a/* in Helsinki corpus from 1350-1710 from Berg & Aronoff (2017)



Simulation of the evolution of multiple spellings for the same affix in a multiagent system: work in progress