Simulating the Early Evolution of Writing

Richard Sproat
Google, Inc

Signs of Writing
Beijing
June 25-27

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

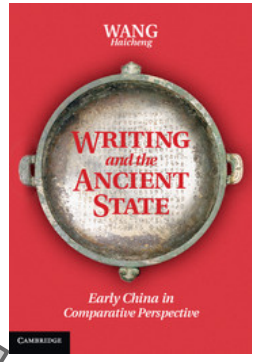
- Computational simulation has been used in a number of areas of linguistics to model phenomena that are hard to test in the laboratory or in the field:
  - Spread of linguistic features in social networks (e.g. Steels, 2012)
  - Historical change (e.g. Niyogi, 2006)
  - Emergence of linguistic properties (e.g. Kirby, 1999 and much subsequent work)
- *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”.
- 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).



Summary of Chicago Presentation

Goals

- Generate a writing system for a language from a non-linguistic symbol system. Symbols are assigned to morphemes, spreading to new morphemes by:
  - Semantic similarity
  - Phonetic similarity
- Determine how many spellings for morphemes are purely semantic, purely phonetic, or mixed
  - Compare these with the situation in true ancient writing systems
- Determine which linguistic properties make phonetic spread easiest.

Parameters: phonotactics

Basic phonotactics for morphemes

Monosyllabic language	σ = C? V R? C?	[R= sonorant]
Disyllabic language	σσ?	

C = p, t, k, b, d, g

V = a, e, i, o, u

R = N, r

Randomly generate ≈1K morphemes from these templates

Parameters: what's phonetically close?

- CLOSE:
  - obstruents match on place of articulation
  - vowels must match
  - sonorants optional
  - **Ex: *daNg* matches *tak***
- CLOSE_RHYME:
  - as above, but "rhyming" is sufficient
  - **Ex: *daNg* matches *bak***
- CLOSE_V_FREE:
  - same as CLOSE, but vowels don't need to match, and V matches ∅
  - **Ex: *donek* matches *dink***
- STRICT:
  - Exact match required

Parameters: semantics

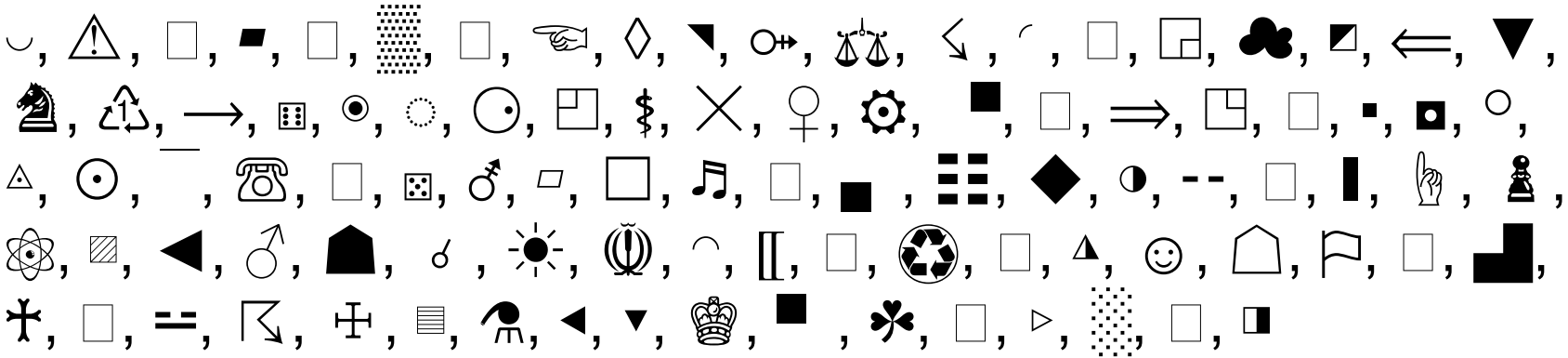
100 basic “concepts”:

PERSON, MAN, WOMAN, HOUSE, BRONZE, SWORD, MEAT, SHEEP, OX, GOAT, FISH, TREE, BARLEY, WHEAT, WATER, STONE, CLAY, THREAD, 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

Randomly associate these with “morphemes” and for remaining morphemes associate those with **random combinations of up to 3 of these**

Parameters: symbols

Each basic concept is associated with a symbol. We used dingbats and other mostly non-script symbols from the Unicode Basic Multilingual Plane.* Some examples:



*Expanded in the current system to include other sets such as Yi symbols

Parameters: “languages”

- Initial parameters:
  - MONOSYLLABIC / DISYLLABIC
  - CLOSE / CLOSE_RHYME / CLOSE_V_FREE / STRICT
- For each pair of the above
  - Generate a “language” 5 times: a language consists of about 1,000 morphemes generated from the grammar
  - Run 5 simulations for each of these

Steps in simulation

1. Assign simplex concepts and symbols to morphemes
   a. **From a set of morphemes associated with a symbol, pick *one* morpheme to associate to the symbol**
2. Assign complex concepts to remaining morphemes
3. Assign spellings to morphemes that do not have one
   a. Symbols associated with shared semantic components, and/or
   b. **Symbols associated with similar sound ⇐ *"eureka" moment***
4. Extend the phonetic coverage with *telescoping*:
   a. *ba* + *ad* → *bad*
5. Result is a mix of semantic and phonetic components

MONOSYLLABIC, CLOSE

toN	PERSON	S(emantic){🏵}
kerg	PERSON	S{🏵}+P(honetic){☆}
gib	PERSON	S{🏵}+P{□}
kak	MAN	S{ ♂ }
gab	WOMAN	S{▫}
pet	WOMAN	S{▫}+P{◺}
urb	HOUSE	S{▲}
kuNt	HOUSE	S{▲}+P{◷}
a	BRONZE	S{☞}
dep	BRONZE	S{☞}+P{☞}
keg	BRONZE	S{☞}+P{☆}

Results from previous simulations

	# entries	w/ spelling	w/out spelling	phon	sem+phon	purephon	sem
DI, CLOSE	661	462 (0.70)	198 (0.30)	47 (0.07)	35 (0.05)	11 (0.02)	451 (0.68)
DI, CLOSE-RHYME	648	508 (0.79)	139 (0.21)	127 (0.20)	100 (0.15)	26 (0.04)	482 (0.74)
DI, CLOSE-V-FREE	644	640 (0.99)	3 (0.01)	324 (0.50)	262 (0.41)	61 (0.10)	579 (0.90)
DI, STRICT-V-FREE	649	504 (0.78)	144 (0.22)	112 (0.17)	87 (0.14)	24 (0.04)	481 (0.74)
MONO, CLOSE	643	580 (0.90)	62 (0.10)	239 (0.37)	193 (0.30)	45 (0.07)	534 (0.83)
MONO, CLOSE-RHYME	651	630 (0.97)	20 (0.03)	310 (0.48)	253 (0.39)	56 (0.09)	574 (0.88)
MONO, STRICT	640	487 (0.76)	152 (0.24)	95 (0.15)	71 (0.11)	23 (0.04)	464 (0.73)

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%)	
Simple indicative	20 (2%)	125 (1%)	107 (1%)	±1,500 (3%)
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

Take away points

- Strict matches are not very effective.
- Close consonant matches useful for largely “monosyllabic” languages; less so for “disyllabic” languages.
- Allowing vowels to be free is useful in “disyllabic” languages.
- The rate of semantic-phonetic compounds in “monosyllabic” languages with close consonant matches is broadly similar to the rate of semantic-phonetic compounds in Oracle Bone texts.
- But we don’t really provide much evaluation of how well the system works

Note: some details of the simulation

- Currently about 1100 lines of Python code
- Uses the *pyfst* interface to OpenFst
- Grammars written in *Thrax* finite-state grammar compiler (http://www.openfst.org/twiki/bin/view/GRM/Thrax)

Evaluating the Similarity to Known Systems

Problems with the previous work

- Letting a symbol inherit the phonology from just *one* morpheme is not a good model of how some ancient systems apparently worked.
- We need some way of quantifying how good a fit the models are to (the few) documented cases of *ex nihilo* evolution of writing.

Similarity within a phonetic class

- For a set of ancient writing systems, we want to measure how phonologically similar the morphemes are within a phonetic class
- For example in Chinese, those morphemes that share the same phonetic radical

Ancient Chinese

- Source: Baxter-Sagart list. (http://en.wiktionary.org/wiki/Appendix:Baxter-Sagart_Old_Chinese_reconstruction)
- Intersect with characters with semantic-phonetic decomposition from http://www.zhongwen.com/.
- Results in 902 entries. Some examples:

丘	丘	qiu1	khjuw	kʷʰə
蚯	丘	qiu1	khjuw	kʷʰə
虛	丘	xu1	khjo	qʰa
虛	丘	xu1	xjo	qʰa
岳	丘	yue4	ngæwk	ŋˁrok

Edit distance (ED)

	a	b	c	d
c	a		e	d

$$\text{Normalized Edit Distance} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\frac{\text{Length}_1 + \text{Length}_2}{2}}$$

- Efficient to compute: $\mathcal{O}(\text{Length}_1 \times \text{Length}_2)$
- Operations could be weighted by phonetic distance.

Ancient Chinese

- 156 equivalence classes (out of 198 — 79%) with >1 member
  - qʷək ɢʷək ɢʷəʔ
  - gˤaŋ gˤraŋ
  - muʔ tʰaʔ tʰˤaʔ tˤaʔ
  - ləp mˤroŋ rˤoŋ
- Compute mean NED for each equivalence class with >1 member; average over all equivalence classes. Smaller is better.

	Mean edit distance
Ancient Chinese	0.54
Middle Chinese	0.60
Modern Mandarin	0.57

Sumerian

- Source: Electronic Text Corpus of Sumerian Literature sign list (http://etcsl.orinst.ox.ac.uk/edition2/signlist.php)

Sign names	Signs	ETCSL values
A		a, dur_5, $duru_5$
A.AN		am_3, em_x, $še\hat{g}_3$
A.EDIN.LAL		ummud
A.HA.TAR.DU		$girim_3$
A.IGI		er_2, $še_x$
A.KA		ugu_2
A.KAL		illu

- Kept all lower-case (phonetic) values, without subscripts.

Sumerian

- 212 equivalence classes (out of 617 — 34%) with >1 member
- Same procedure as with Chinese

	Mean edit distance
Ancient Chinese	0.54
Middle Chinese	0.60
Modern Mandarin	0.57
Sumerian	0.89

Middle Egyptian

- 21 equivalence classes (out of 256 — 8%) with more than one member from Gardiner's sign list (via http://wwwegyptianhieroglyphsnet/gardiners-sign-list/low-narrow-signs/),

	Mean edit distance
Ancient Chinese	0.54
Middle Chinese	0.60
Modern Mandarin	0.57
Middle Egyptian	0.60
Sumerian	0.89

Two neurological routes to grammatogenesis

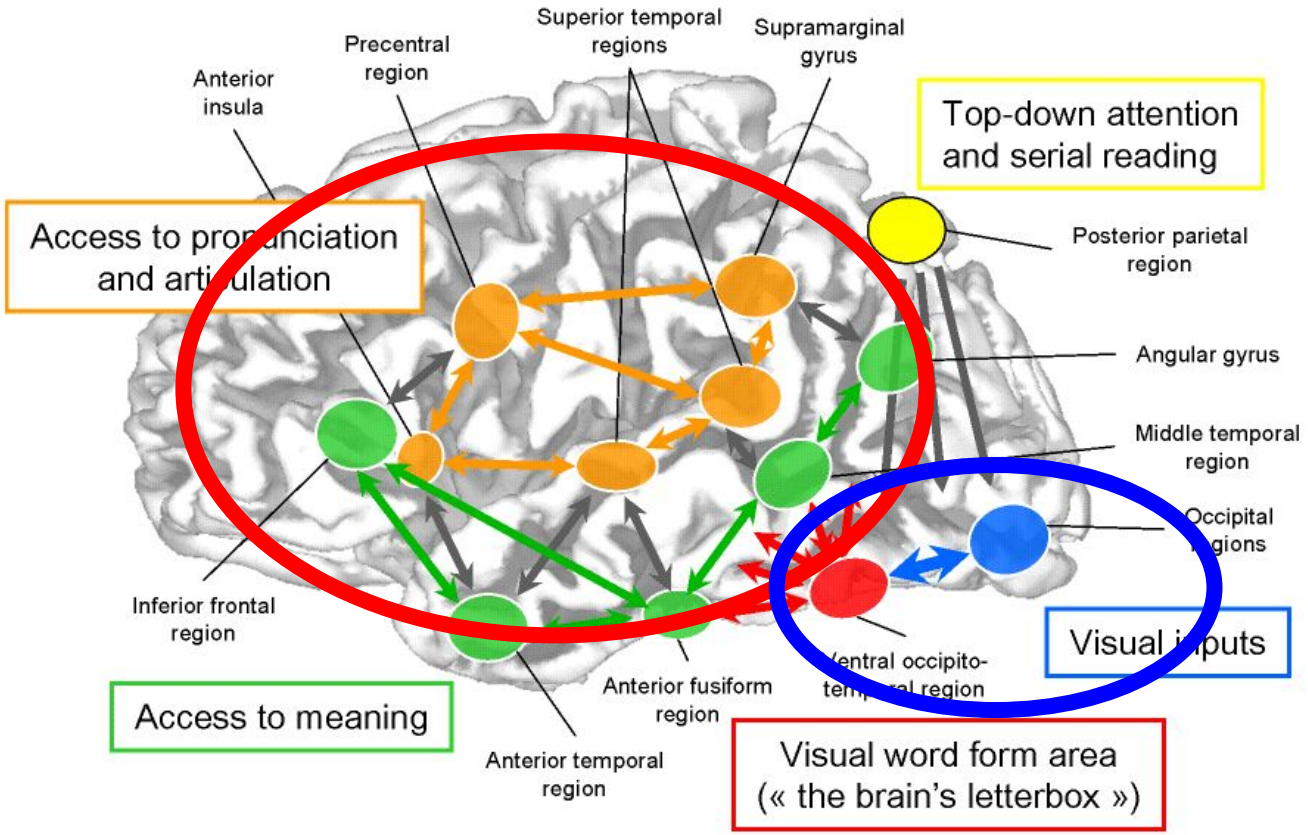


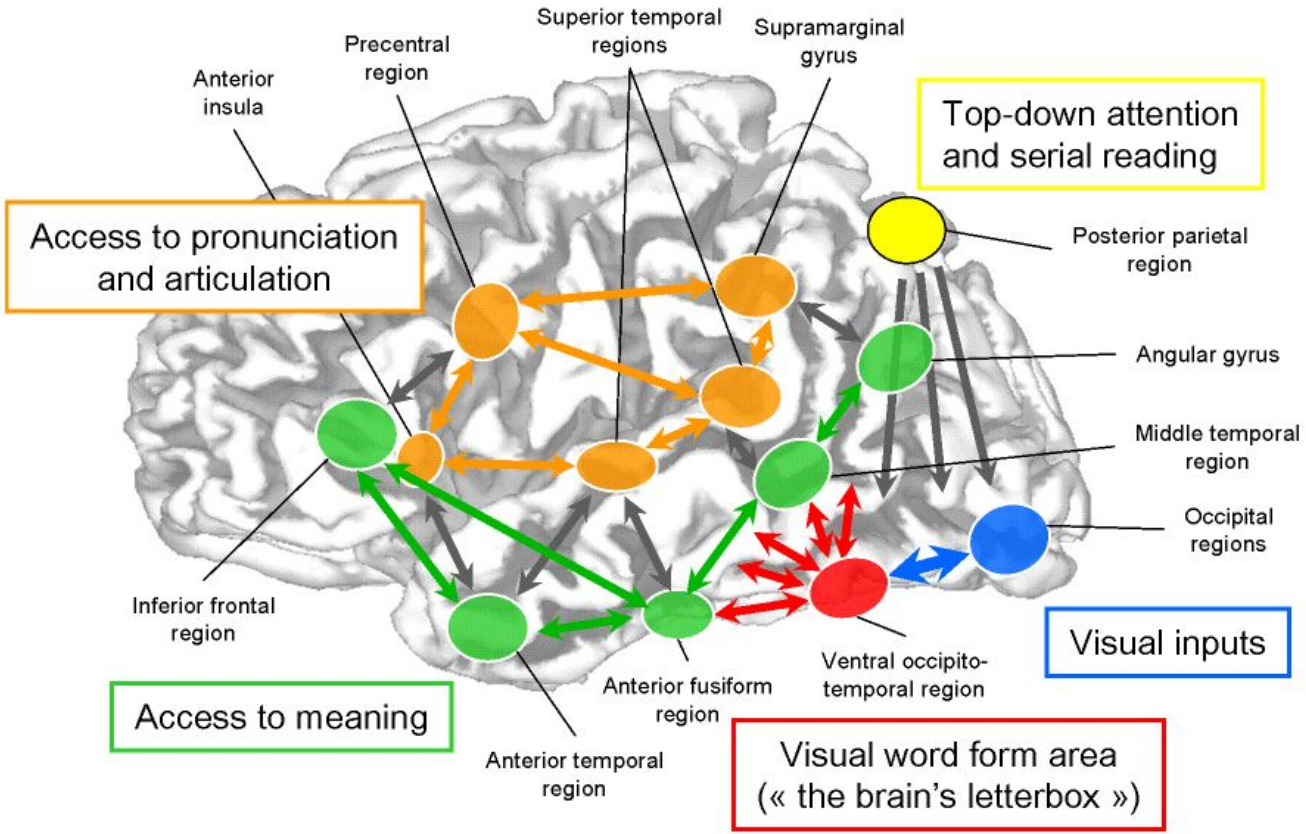
Image from:

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

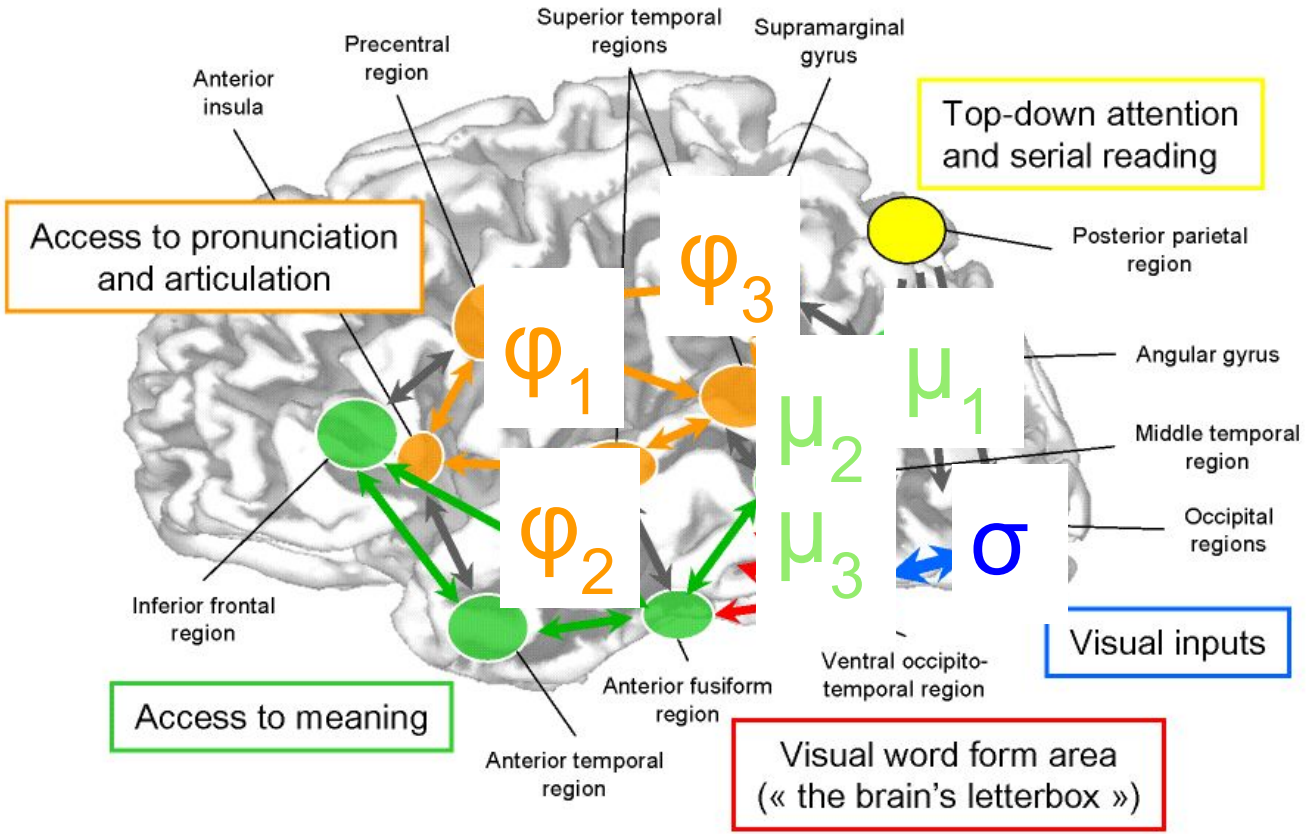
Language and speech

Visual processing

Two neurological routes to grammatogenesis



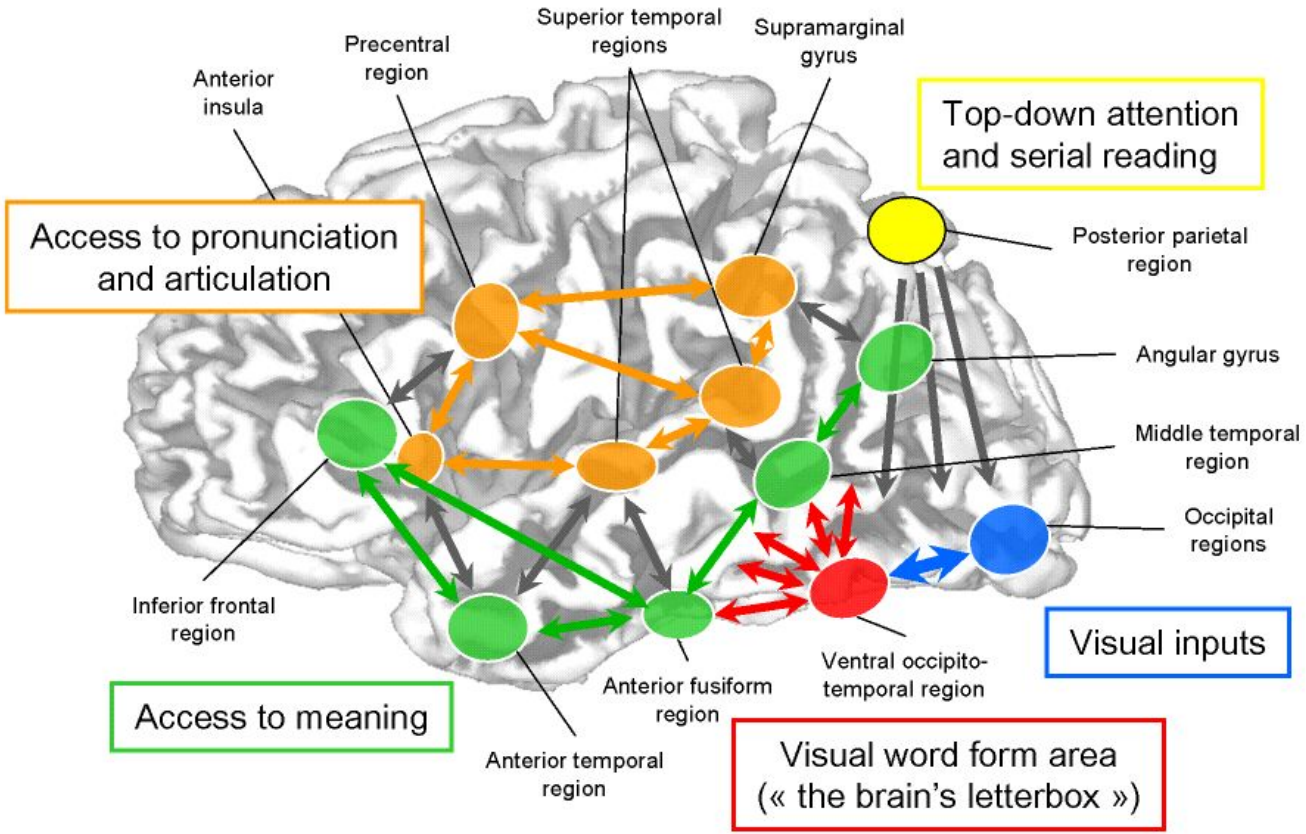
Two neurological routes to grammatogenesis



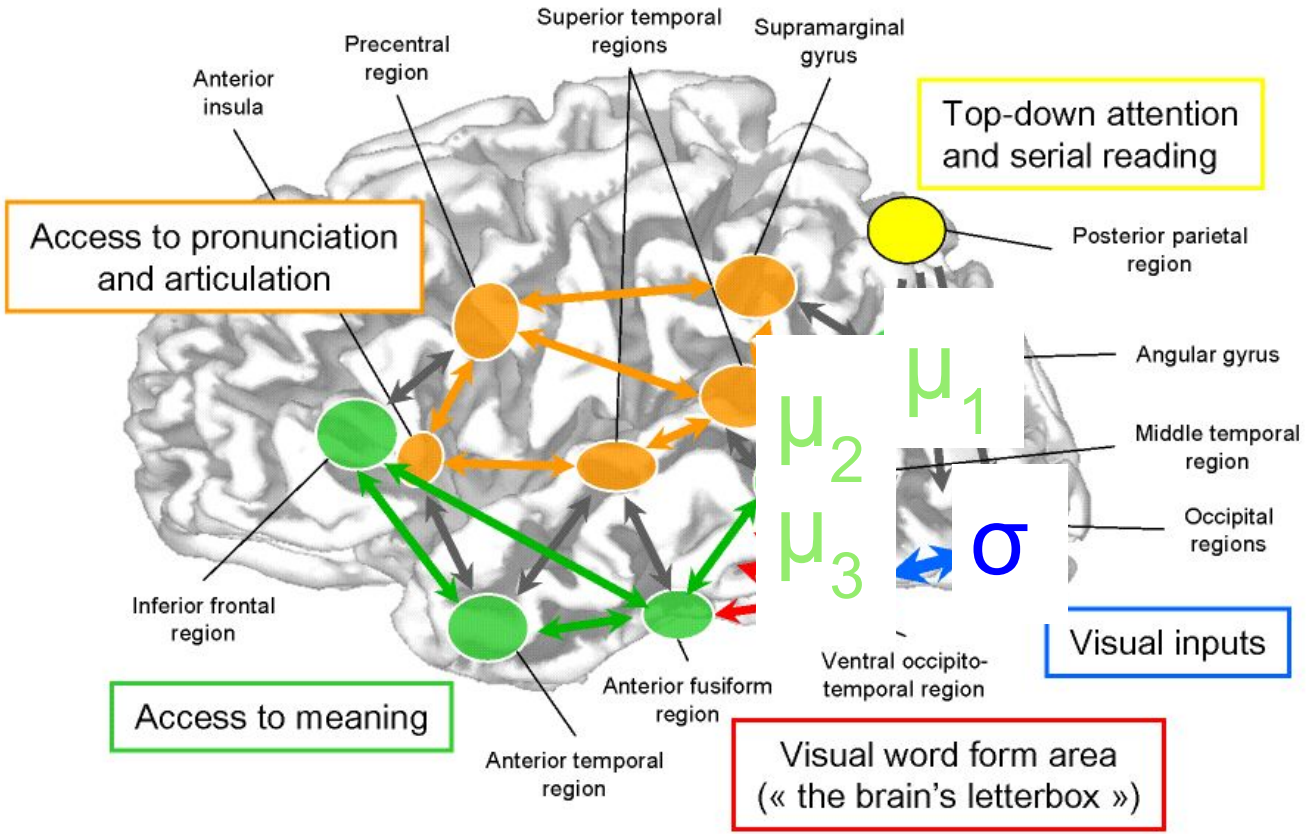
Symbol σ associated with concept and thence a set of related meanings/morphemes μ_k

Via the morphemes becomes associated to set of pronunciations φ_k

Two neurological routes to grammatogenesis



Two neurological routes to grammatogenesis

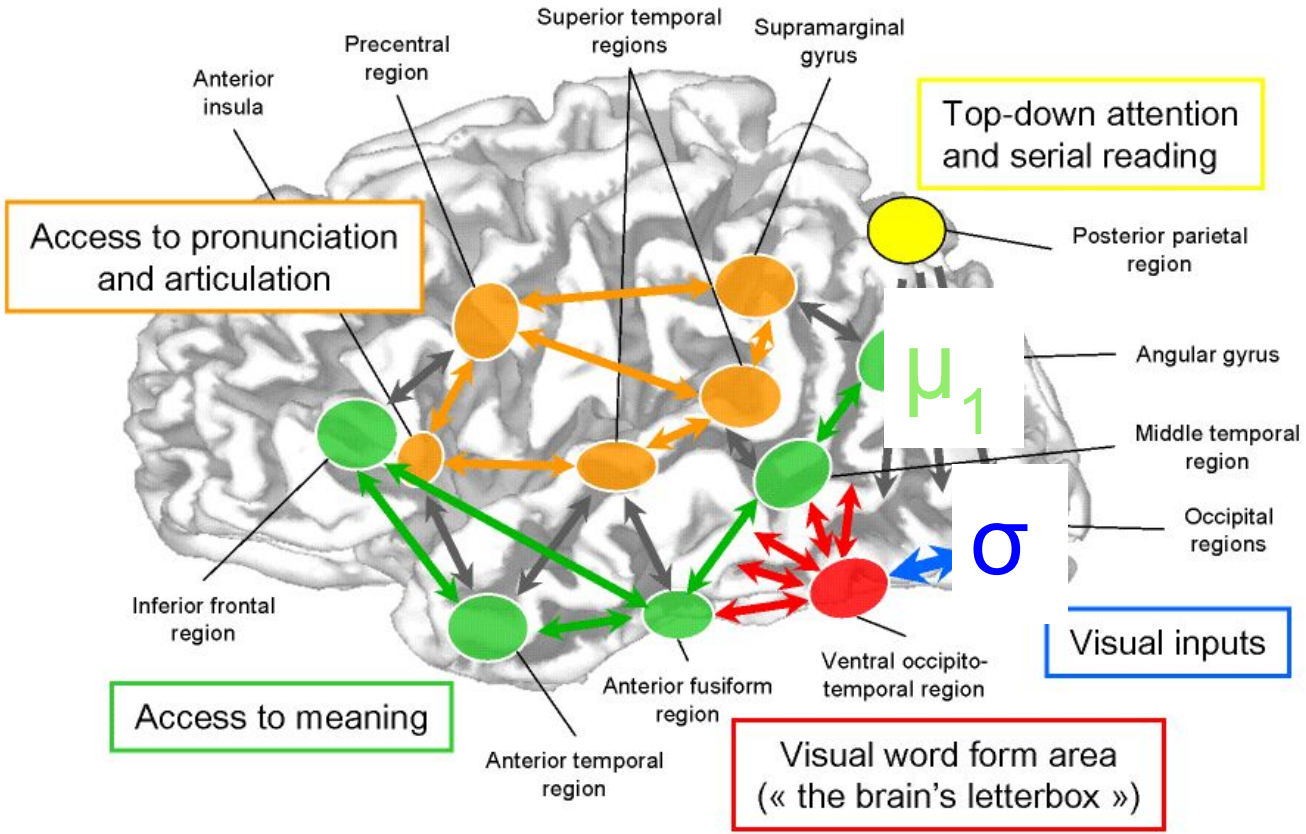


Symbol σ associated with concept and thence a set of related meanings/morphemes μ_k

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

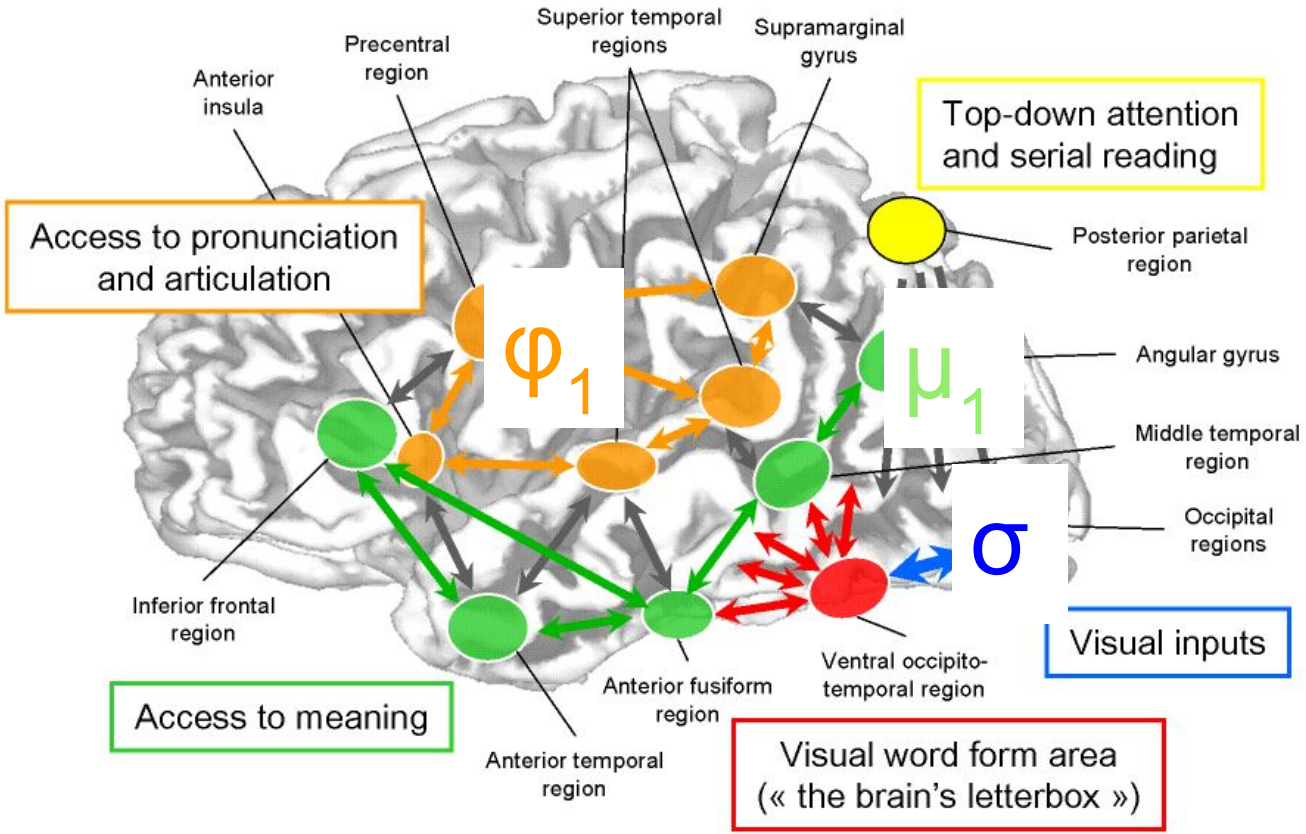


Symbol σ associated with concept and thence a set of related meanings/morphemes μ_k

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



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.*

Simulation

Parameterize with two options:

- As in previous simulations, pick just one morpheme per semantic group and base phonetics on that morpheme (Chinese, Egyptian)
- Use all morphemes in a semantic group (Sumerian)

Coverage results

	# entries	w/ spelling	w/out spelling	phon	sem+phon	purephon	sem
DI, CLOSE	661	462 (0.70)	198 (0.30)	47 (0.07)	35 (0.05)	11 (0.02)	451 (0.68)
DI, CLOSE-RHYME	648	508 (0.79)	139 (0.21)	127 (0.20)	100 (0.15)	26 (0.04)	482 (0.74)
DI, CLOSE-V-FREE	644	640 (0.99)	3 (0.01)	324 (0.50)	262 (0.41)	61 (0.10)	579 (0.90)
DI, STRICT-V-FREE	649	504 (0.78)	144 (0.22)	112 (0.17)	87 (0.14)	24 (0.04)	481 (0.74)
MONO, CLOSE	643	580 (0.90)	62 (0.10)	239 (0.37)	193 (0.30)	45 (0.07)	534 (0.83)
MONO, CLOSE-RHYME	651	630 (0.97)	20 (0.03)	310 (0.48)	253 (0.39)	56 (0.09)	574 (0.88)
MONO, STRICT	640	487 (0.76)	152 (0.24)	95 (0.15)	71 (0.11)	23 (0.04)	464 (0.73)

	# entries	w/ spell	w/out sp	phon	sem+phon	purephon	sem
MONO CLOSE 1 morph	638	587 (0.92)	**50 (0.08)****	**235 (0.40)****	**203 (0.32)***	**50 (0.08)****	**537 (0.84)****
MONO CLOSE all morphs	645	636 (0.99)	**8 (0.01)****	**222 (0.34)****	**187 (0.29)***	**34 (0.05)****	**601 (0.93)****

Comparison with extant systems

	Mean edit distance	Standard deviation
Ancient Chinese	0.54	
Middle Chinese	0.60	
Modern Mandarin	0.57	
Middle Egyptian	0.60	
MONOSYLLABIC_CLOSE: 1 morph*	0.52	0.026
MONOSYLLABIC_CLOSE: all morphs*	0.78	0.026
Sumerian	0.89	

*In both cases the percentage of signs with multiple pronunciations is about 64% — less than Chinese but more than Sumerian or Egyptian

Simulation using OC syllables

- 404 Old Chinese reconstructed syllables:
  - *ba, baŋ, ben, but, bˁa, bˁawk, bˁaŋ, bˁrak…*
- 198 equivalence classes used to define “CLOSE”
- One morpheme per group

Example: ☘

lrə LEAF S{☘}

lek LEAF,BIRTH S{☘,㋮}

dʕak CUP,LEAF S{☘}+P{⚜}

paʔ NIGHT,HOUSE,LEAF S{☘, ☾ }

dʕək DEATH,FIRE,LEAF S{☠,☲,☘}

lek NIGHT,HOUSE,LEAF S{☘, ☾ ,☖}

lʕəʔ ENCLOSURE,SKIN,EYE S{⚲ ,◎}+P{☘}

lə GOD,DEMON,OX S{ ☿ ,♉}+P{☘}

Example: ☽

n̥ij	MOON S{ ☽ }
mət	THUNDER,FIRE,MOON S{䷳, ☽ }
rˁew	WOMB,ARM,MOON S{ ☽ , ⚕}
nij	COAL S{■}+P{ ☽ }
nij	CLOTHING,SUMMER,THUNDER S{⚡}+P{ ☽ }
nij	SILVER,FISH S{𓆛 }+P{ ☽ }

Coverage results

	# entries	w/ spelling	w/out spelling	phon	sem+phon	purephon	sem
DI, CLOSE	661	462 (0.70)	198 (0.30)	47 (0.07)	35 (0.05)	11 (0.02)	451 (0.68)
DI, CLOSE-RHYME	648	508 (0.79)	139 (0.21)	127 (0.20)	100 (0.15)	26 (0.04)	482 (0.74)
DI, CLOSE-V-FREE	644	640 (0.99)	3 (0.01)	324 (0.50)	262 (0.41)	61 (0.10)	579 (0.90)
DI, STRICT-V-FREE	649	504 (0.78)	144 (0.22)	112 (0.17)	87 (0.14)	24 (0.04)	481 (0.74)
MONO, CLOSE	643	580 (0.90)	62 (0.10)	239 (0.37)	193 (0.30)	45 (0.07)	534 (0.83)
MONO, CLOSE-RHYME	651	630 (0.97)	20 (0.03)	310 (0.48)	253 (0.39)	56 (0.09)	574 (0.88)
MONO, STRICT	640	487 (0.76)	152 (0.24)	95 (0.15)	71 (0.11)	23 (0.04)	464 (0.73)

	# entries	w/ spell	w/out sp	phon	sem+phon	purephon	sem
MONO CLOSE 1 morph	638	587 (0.92)	50 (0.08)	235 (0.4)	203 (0.32)	50 (0.08)	537 (0.84)
MONO CLOSE all morphs	645	636 (0.99)	8 (0.01)	222 (0.34)	187 (0.29)	34 (0.05)	601 (0.93)
“Ancient Chinese” 1 morph	650	533 (0.82)	116 (0.25)	163 (0.25)	130 (0.20)	32 (0.05)	500 (0.77)

Comparison with simulations

	Mean edit distance	Standard deviation
Ancient Chinese	0.54	
Middle Chinese	0.60	
Modern Mandarin	0.57	
Middle Egyptian	0.60	
MONOSYLLABIC_CLOSE: 1 morph	0.52	0.03
“Ancient Chinese”-based: 1 morph	0.67	0.08
MONOSYLLABIC_CLOSE: all morphs	0.78	0.03
Sumerian	0.89	

Synopsis and conclusions

- A model where pronunciations spread from all morphemes associated with a concept fits Sumerian better.
- A model where one morpheme is picked as *the* denotation of the symbol fits Chinese (or Egyptian) better.
- Modeling based on “actual” Ancient Chinese syllables and limiting ourselves to reconstructed phonetic equivalence classes does not work well — but that is probably because the phonetic equivalence classes are too limiting.
- Proportion of signs with multiple pronunciations also differs significantly from known systems. This could partly reflect standardization in real systems, which we are not modeling.

Synopsis and conclusions

- 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...

Further work

- Link consonantal systems to ablaut-like processes
  - *We already have some results but no time to report here.*
- Simulate writing full sentences and affixal morphology
  - *We already have some results but no time to report here.*
- Simulate a wider range of phonetic shapes for morphemes, and a wider range of phonetic closeness
- Provide a more plausible model of lexical statistics
- Model of standardization processes
- Set of symbols is currently static: but new symbols are invented in real writing systems