Glyph structure

A prosodic/segmental/featural framework

James Myers
National Chung Cheng University
Taiwan

https://lngmyers.ccu.edu.tw/

Grapholinguistics in the 21st century - Reading 2026

Thanks!

- National Science and Technology Council (Taiwan) grant 112-2410-H-194-030-MY3
- My chief lab assistant Yangru Chen
- Anonymous reviewers, Lian-Hee Wee
- YOU! (I hope)

Overview

- Glyphs (basic graphetic shapes) are formally systematic, but how?
- No single explanatory tool seems to be enough on its own
- So we need a **multilevel framework**
  - Overall structure vs. stroke inventories vs. stroke properties
  - Analogous to spoken/signed phonology: Prosody vs. segments vs. features
- This is an old idea but it deserves more systematic exploration
  - How should each of these levels be motivated and formalized?
  - Does the framework really make sense for all writing systems?

Glyphs

- Glyphs are written basic shapes (Meletis & Dürscheid 2022)
  - Latin letters: C, H (cf. CH)
  - Chinese character components: 女 *nǚ* ‘woman’, 子 *zǐ* ‘child’ (cf. 好 *hǎo* ‘good’)
  - Simple Devanagari akshara: क *ka*, ल *la* (cf. क्ल *kla*) (Gnanadesikan 2022)
  - Yi syllabograms: *bit*,  *pit*, *bi*,  *pi*,  *bot*, *pot*
- 99.99% of writing research is on glyph combinations & their mapping to spoken language (graphematics)
- But glyphs themselves are formally systematic, with inventories showing clear coherence (AKA combinatoriality: Kim et al. 2025)
  - Incoherent glyph inventory: C, 女, क,
  - Cf. non-glyph inventories:    

Global glyph features? (1)

- Objective quantification of overall glyph complexity
  - Altmann (2004), Peust (2006), Nag et al. (2014), Miton & Morin (2021), ...
  - But quantification fails to capture the discrete nature of glyph contrasts

    P vs. B vs. R (cf. - vs. – vs. —)
- Subjective features for overall glyph form contrasts
  - Gibson (1969): perception task on blurry uppercase Latin letters → confusion matrix (R misread as B etc) → glyph-level binary features
  - Fiset et al. (2008): similar task though different in detail → a totally different set of glyph-level binary features
  - Kim et al. (2025): participants view glyphs in unfamiliar writing systems → asked to propose glyph-level binary features → separate feature sets for every writing system

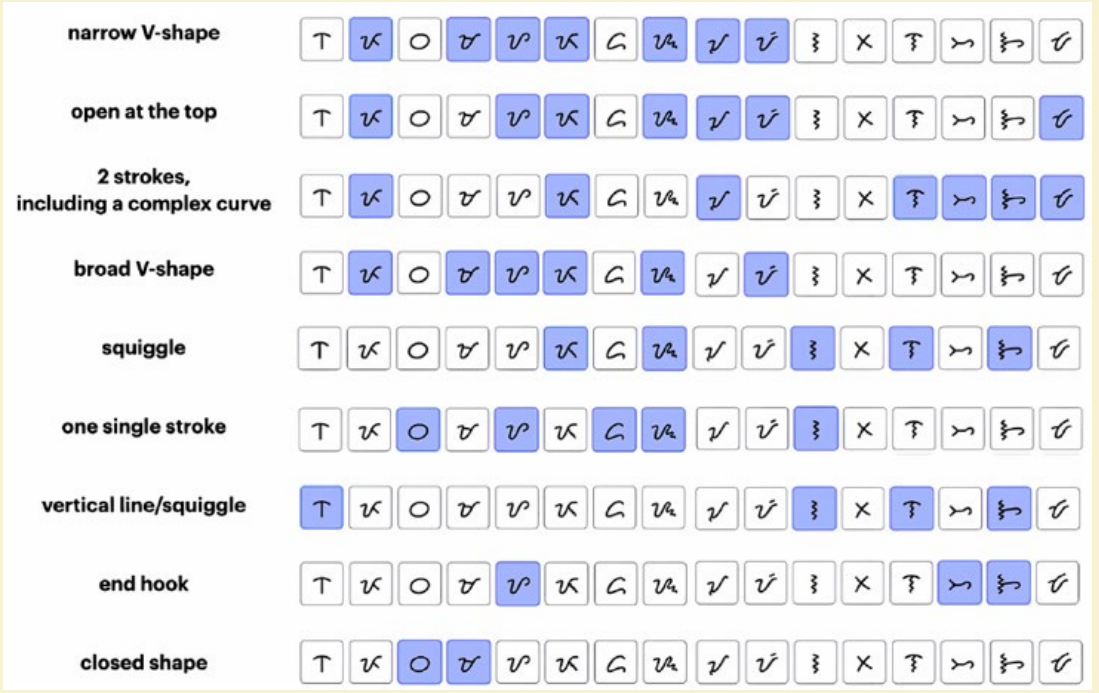
Global glyph features? (2)

- Gibson features for Latin letters

Features	A	E	F	H	I	L	T	K	M	N	V	W	X	Y	Z	B	C	D	G	J	O	P	R	Q	S	U
Straight																										
horizontal	+	+	+	+		+	+								+				+							
vertical		+	+	+	+	+	+	+	+	+				+		+		+				+	+			
diagonal /	+							+	+		+	+	+	+	+											
diagonal \	+							+	+	+	+	+	+	+									+	+		
Curve																										
closed																+		+			+	+	+	+		
open V																				+						+
open H																	+		+	+					+	
Intersection	+	+	+	+			+	+					+			+						+	+	+		
Redundancy																										
cyclic change		+							+			+				+									+	
symmetry	+	+		+	+		+	+	+		+	+	+	+		+	+	+			+					+
Discontinuity																										
vertical	+		+	+	+		+	+	+	+				+								+	+			
horizontal		+	+			+	+								+											

(Gibson 1969)

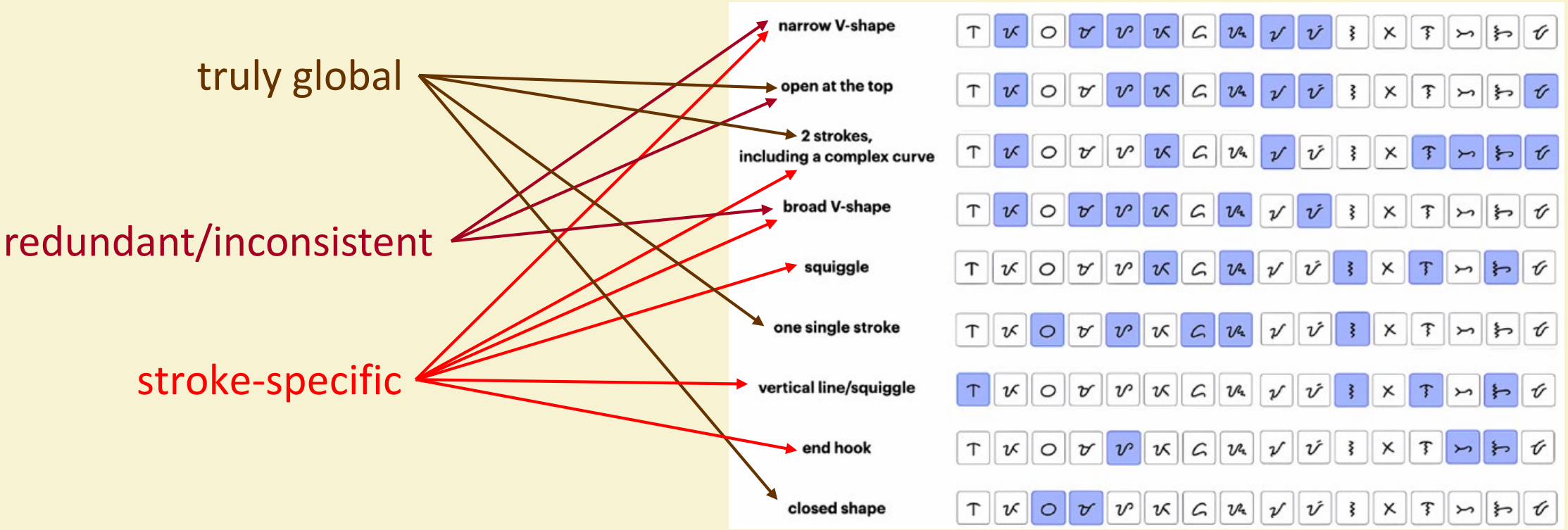
- Kim features for Tagbanwa



(Kim et al. 2025)

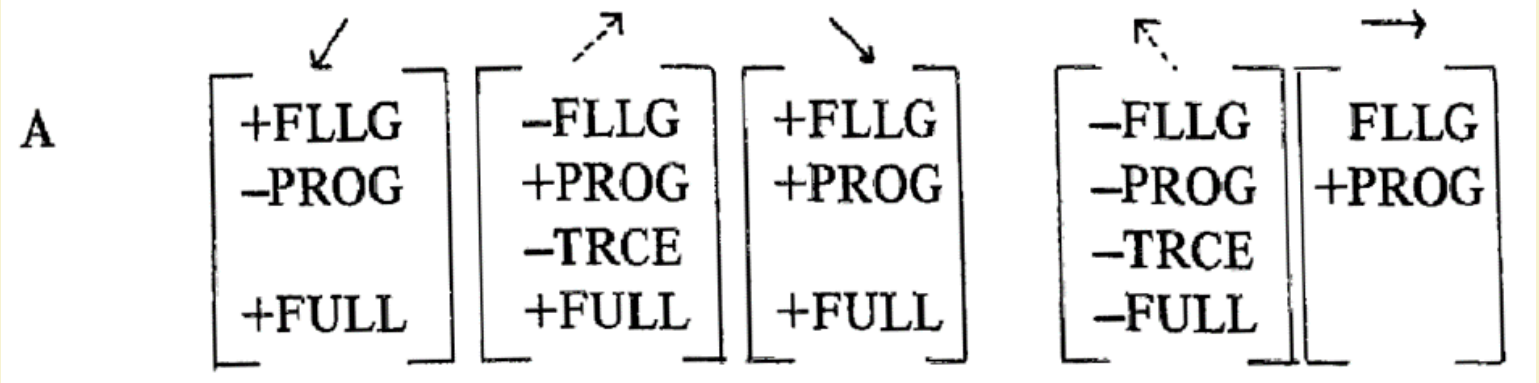
But glyphs are too complex for global features

- Watt's (1980) criticism of Gibson's features (like HORIZONTAL): exactly which part or parts is/are horizontal, and where are they?
- Similar criticisms apply to Fiset and Kim features



Glyphs as combinations of basic elements?

- Glyphs are composed of strokes, which are what contrast in features
  - Latin letters: Eden (1961), Althaus (1973), Watt (1980), Primus (2004)

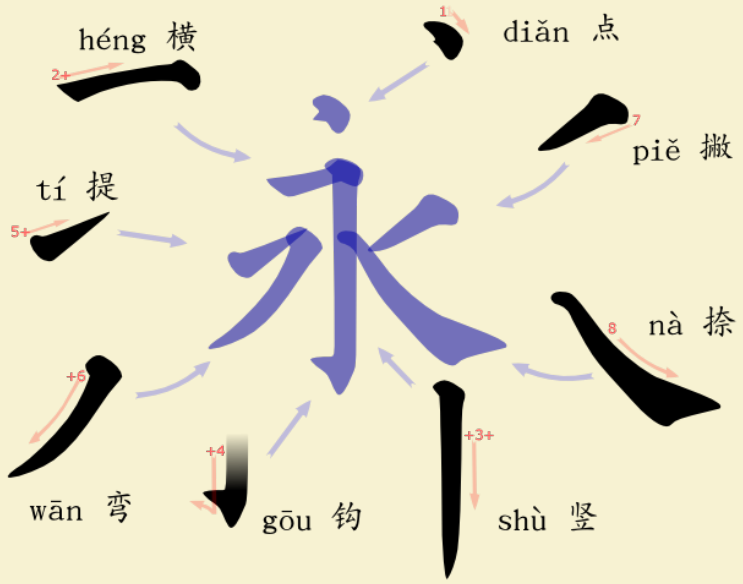


(Watt 1980)

  - Chinese character components: Wang (1983), Peng (2017), Myers (2019)
    - (Next slide please)

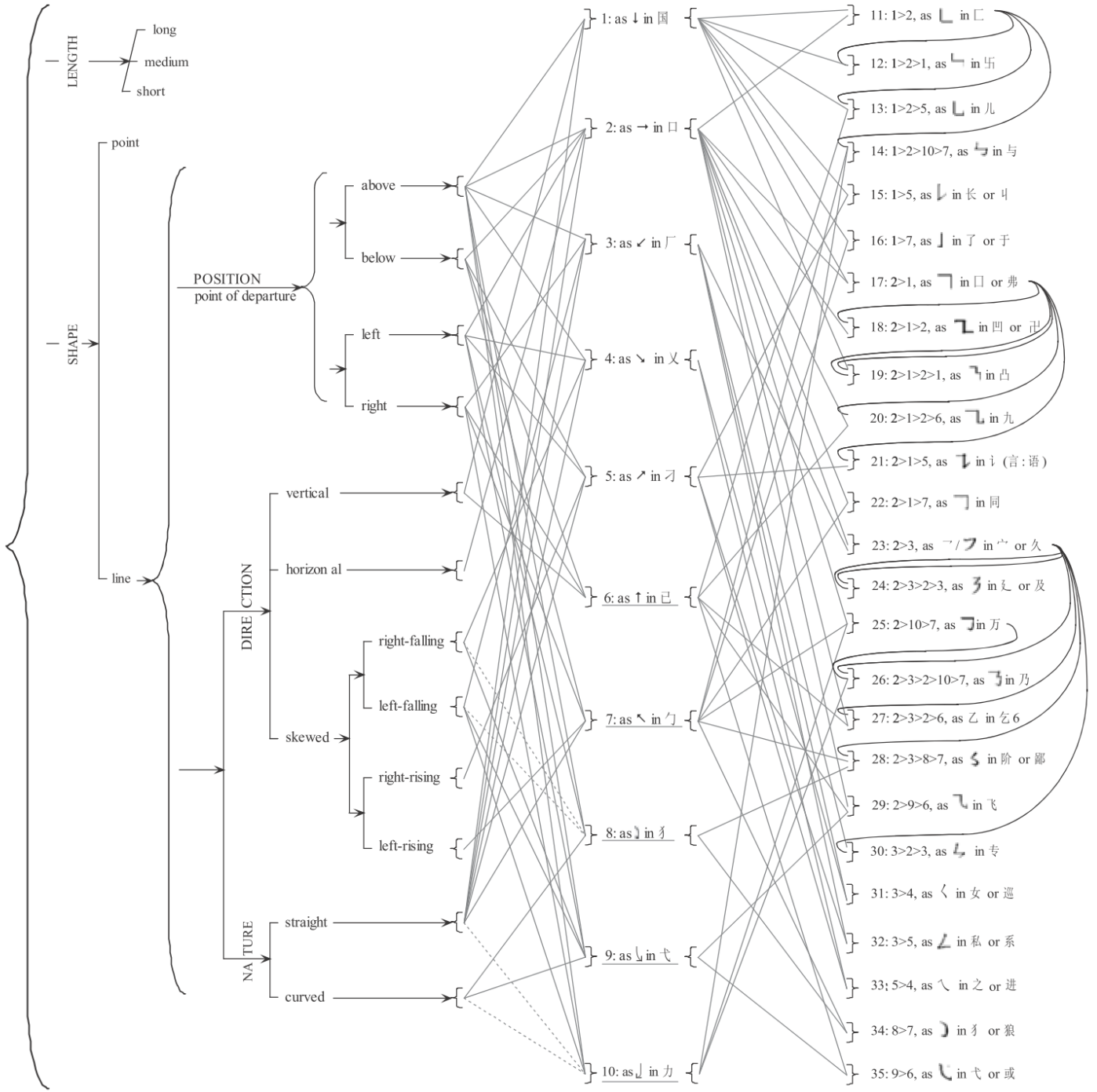
Chinese character strokes and their contrastive features

Wikipedia ↓



https://commons.wikimedia.org/wiki/File:8_strokes_of_%E6%B0%B8-zh.png

Peng 2017 →



Decomposing complex strokes

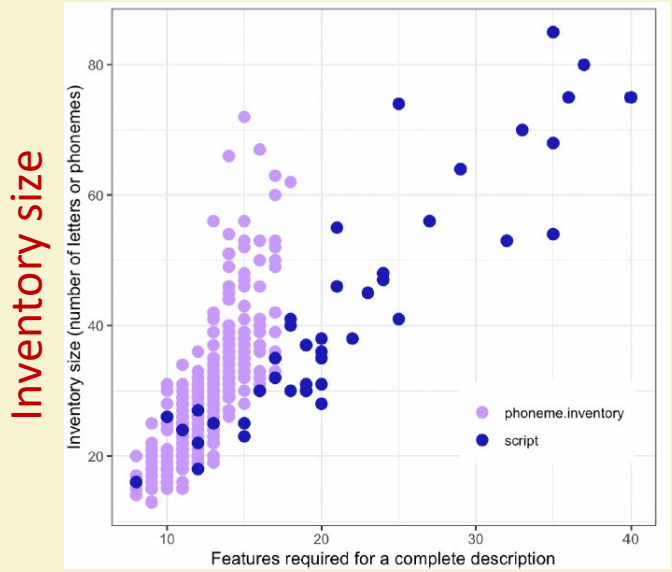
- Complex strokes are composed of substrokes (basic strokes):
  - Already the conventional view in Chinese: 𠃌 = 一 丨 一 亅
- Features only make sense for substrokes
  - Chinese: Curving only in final substroke in 𠃌 = 一 丨 一 ノ
  - Tagbanwa: Wiggling* only in final substroke in ᜐ *sa* (vs. ᜌ *ya*)

    (*A substroke feature, not a complex stroke: cf. trills in spoken & signed phonology)
- Within complex strokes, substrokes interact like separate strokes
  - E (four strokes) vs. L (one stroke), but same | _ interaction at bottom
  - Caveat: Juncture type must be specifiable: Chinese 𠃍 (sharp) vs. 𠃍 (curved)
  - And local constraints can apply: Chinese 乙 = 一𠃍 (also no basic stroke ㄴ)
- Decomposition of circles: an implicational universal?
  - If arcs, then circles: ᑕ ⇒ ᑐ, ᑎ *ṭa* ⇒ ᑏ *ṭha*, ᐯ *pa* ⇒ ᐱ *lip*, hiragana: つ *tsu* ⇒ の *no*
  - Chinese has neither; Korean hangul only has circles

But combinatoriality isn't enough

- Undergeneration: Can't explain stroke allography (Rezec 2009)
  - If E = |- - - , then why is each - different?

- Overgeneration: Can't effectively rule out unattested glyphs
  - Global **glyph** features are far less economical and informative than features that distinguish between **phonemes** (Kim et al. 2025) →
  - Strokes in any given inventory can be combined in ways that go beyond the attested glyphs: L T vs. ⅃ ⊥

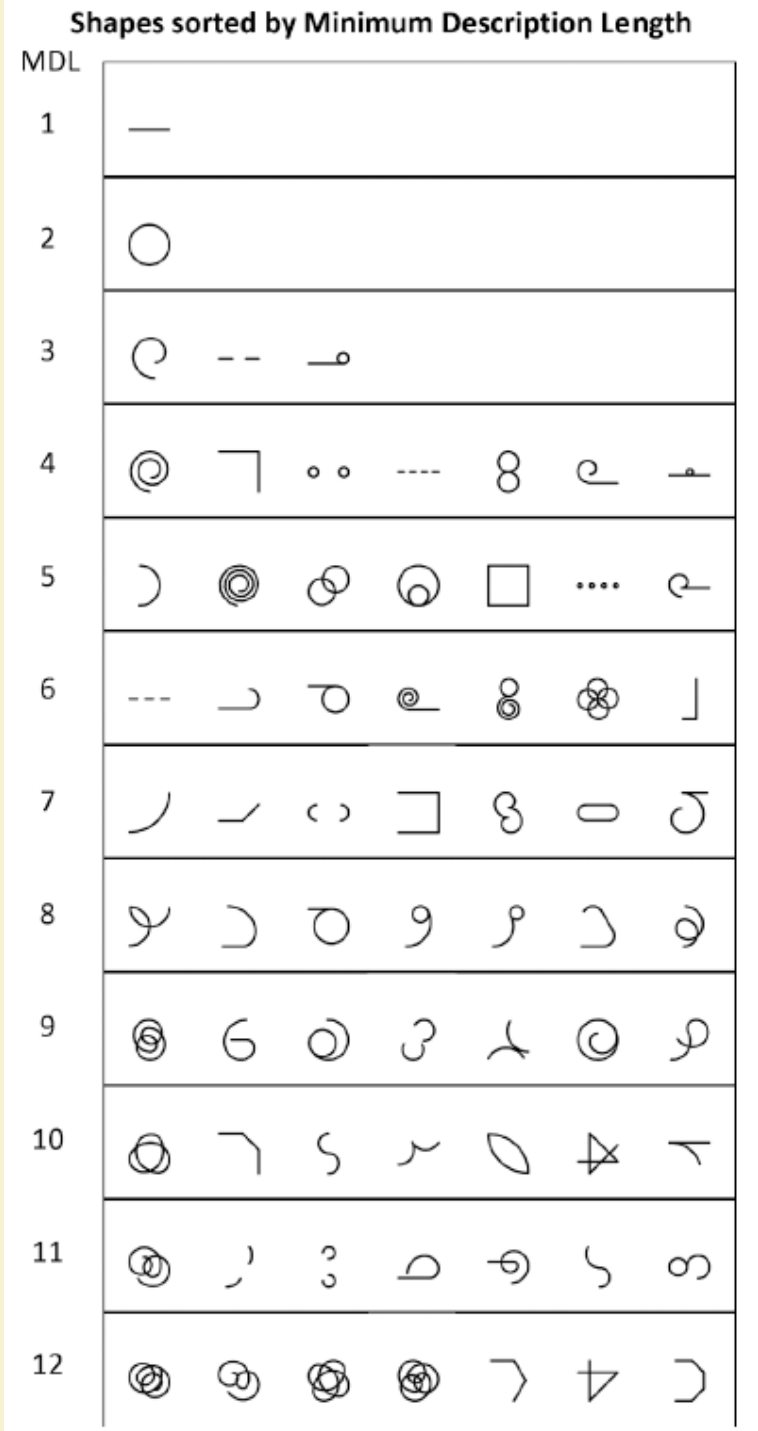


Generative algorithms

- Sablé-Meyer et al. (2022)
  - For human-produced geometric shapes

```
Repeat(next(next(one))) {
Subprogram {Trace(acceleration=-one/next(next(one)),
turningSpeed=one)
};
Turn(angle=next(next(next(one)))/next(next(one)))
}
```

  - But not only don't these look like glyphs...



… but such algorithms always overgenerate

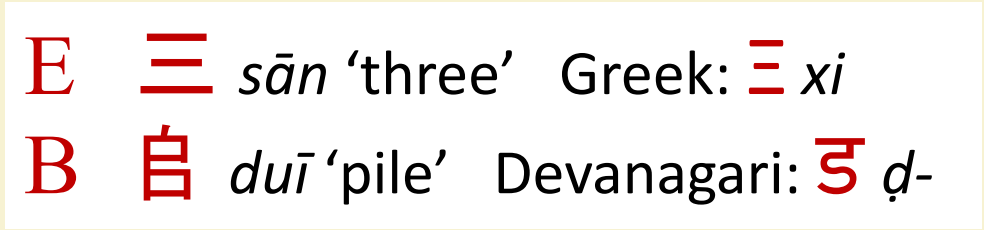
- " 'Excessive power' arguments are arguments that a notational system is wrong because it allows one to express impossible rules. [...] [This] is of course unheard of outside of linguistics – no mathematician criticizes a notation on the ground that it allows one to write the sentence 2 + 2 = 59."

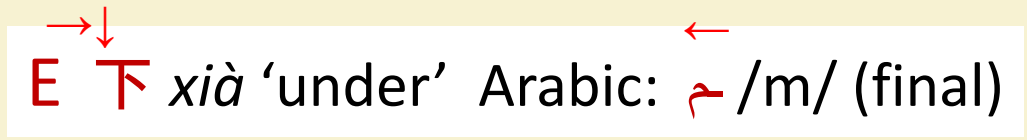
  McCawley (1973 [1979])

- The standard solution: Constraints!

Constraining combinatoriality

- Reducing undergeneration
  - Global position constrains stroke form
    - Bottom stroke most enlarged, top stroke secondarily so (Myers 2019)
  - Local position also constrains stroke form
    - Stroke combinations avoid oblique angles (Morin 2018)
    - Strokes tend to start, not end, at others (van Sommers 1984; Primus & Wagner 2013)



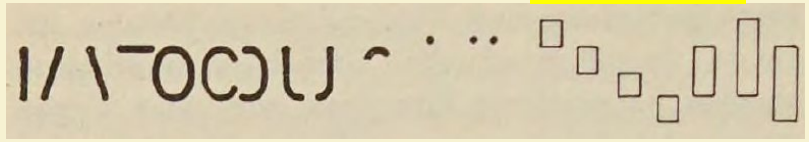


- Reducing overgeneration
  - Unlike features, positional constraints are systematic rather than contrastive
    - Contrasts like 土 (*tǔ* ‘earth’) vs. 士 (*shì* ‘scholar’) are rare (Yang & Wang 2018)

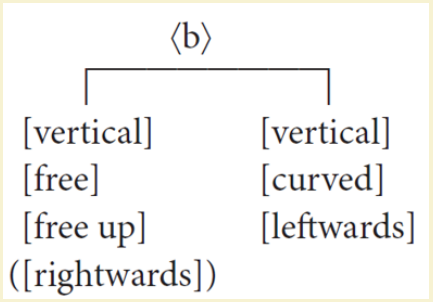
So positions also matter

- Latin letters
  - Althaus (1973)

Elements Positions



  - Primus (2004)

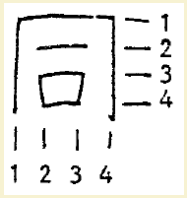


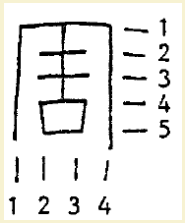
I in head position, **ɔ** in coda position

Features

- Chinese character components
  - Wang (1983): Curve leftmost stroke only in relatively "tall" components

同 *tóng* 'same'
周 *zhōu* 'circle'



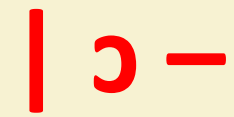


A multilevel approach

Glyphs

Global positional constraints

Local constraints



Strokes

hor, vert, curved, … Stroke features

A multilevel approach ... that's oddly familiar

B 自	Glyphs	**Morphemes***	**story**
O	Global positional constraints	**Stress feet**	**stóry**
X ⊢→	Local constraints	**Syllables & segmental constraints****	**s[t]o.ri**
| ɔ –	Strokes	**Segments**	**s t o r i**
hor, vert, curved, ...	Stroke features	**Segmental features**	**voiceless, coronal, ...**

*(Eden 1961, Watt 1980, Myers 2019, Gnanadesikan 2022, ...)

**(Thanks, anonymous reviewer!)

Why the familiarity is useful

- While multi-module models are harder to test empirically...
- ... we know all these modules already, including the caveats
  - Global morpheme features would be absurd: *story* as [+disyllabic, +s, ...]
  - Prosody is generally not contrastive (Cutler 2015)
    - No minimal stress pairs in English morphemes (*ínsight* vs. *incíte*)
  - Segments can be hard to segment
    - Does *James* have four phonemes /dʒ eɪ m z/ or six /d ʒ e ɪ m z/?
  - Features aren't as universal as once thought (Mielke 2008)
    - /l/ may be a stop or a continuant, depending on the language
    - Spoken vs. signed features are totally different
    - So it's not fatal that stroke inventories differ so much across scripts

“Prosodic” constraints

- **Global** constraints share formal similarities with **metrical prosody**
  - Asymmetry (***stó**ry*)
    - Glyphs tend to “face” one way: **BCDEFGKLPQR** (cf. **J**) /m/ **ध** *dha* **न** *na*
    - And to be bottom-heavy: **三**  **ㄷㅎ** **ABGJLMQR** (cf. **FPTVWY**)
  - Abstract (not featural): bottom-heaviness in modern **木** *mù* ‘tree’ vs. ancient 
  - Binarity (*sto.ry*)
    - **BFKMSWXZ** (cf. **E**) ****(cf. **三**)   *lop* *jot* (cf. *tap*) **घ** *gha*
- Some **local** constraints share formal similarities with **syllables**
  - Syllables favor onsets over codas
    - Easier to coordinate gestural starts than ends (Browman & Goldstein 1988)
  - Similar to the principle of starting strokes at others (Myers 2021)
    - 二 *èr* ‘two’ & 工 *gōng* ‘work’ are both “disyllabic” (一一 & 丅一)

Tricky cases

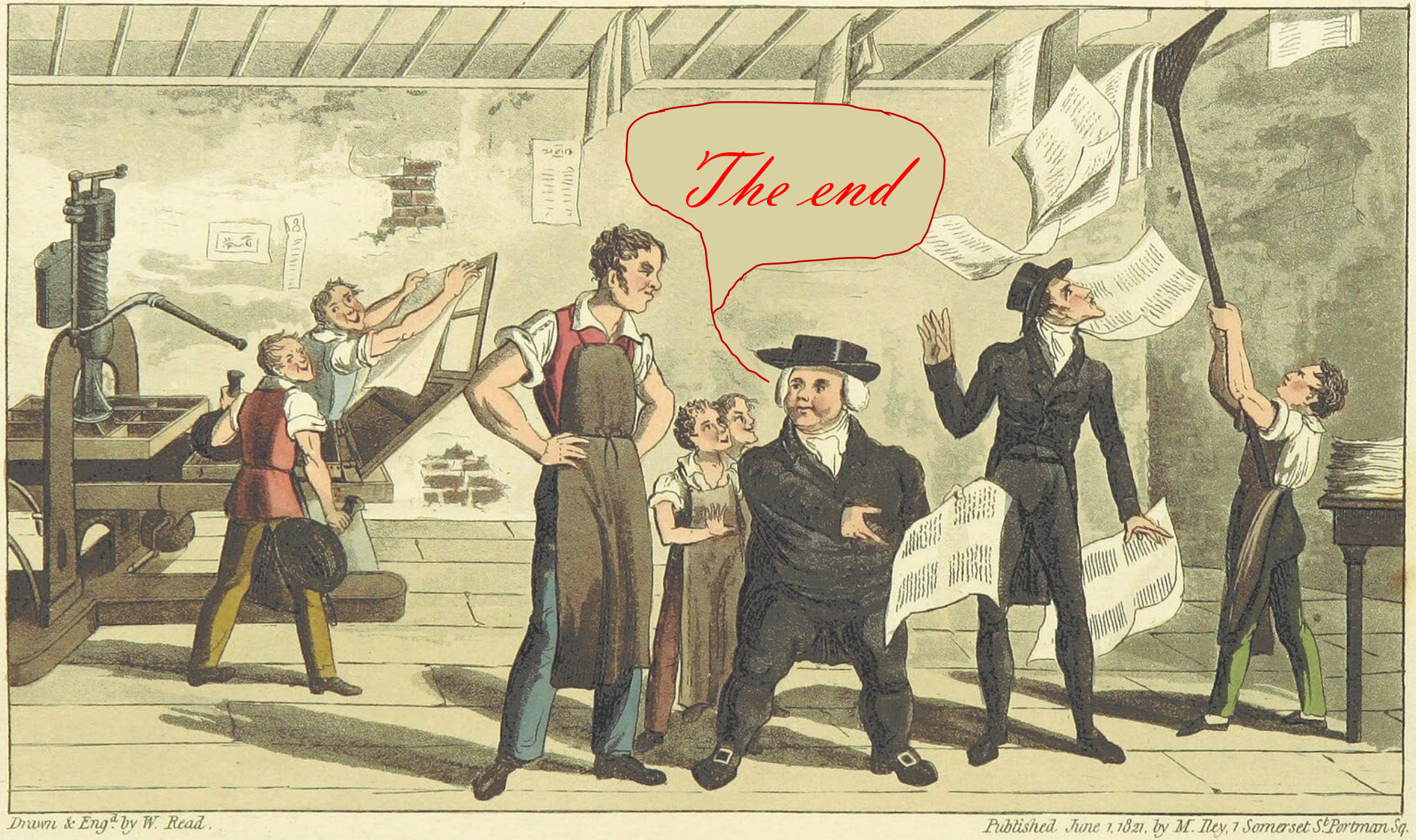
- Egyptian hieroglyphs
  - Emoji-like detail and iconicity:  ‘eat’  ‘embrace’  ‘40,000’
  - Yet it still has global constraints (e.g. facing left):  ‘frog’  ‘bee’
  - And a (very) few formal (?) near-minimal pairs:  ‘great’  ‘bad’*
  - But phonology has to evolve, as in natural sign languages (Sandler et al. 2011)
  - And iconicity can complement formal phonology, again as in natural sign languages (Eccarius & Brentari 2010)
- Chinese Bopomofo (Zhuyin fuhao)
  - Arcs but no circle: ㄛ /o/ ㄜ /ɤ/
  - Because this script didn’t evolve naturally...?

*(Thanks again, anonymous reviewer!)

Conclusions

- A prosodic/segmental/featural framework for glyph structure, analogous to well-established frameworks for speech and signing, may help combine the strengths and escape the limitations of other glyph frameworks
- As a richly detailed framework, it can also help guide future empirical work in typology and psycholinguistics
- Why does this analogy seem to work?
  - Did the biological evolution of language result in a relatively amodal system?
  - Is the cultural evolution of language restricted by an intrinsic solution space?

https://en.wikipedia.org/wiki/Charles_Williams_(caricaturist)#Works
https://commons.wikimedia.org/wiki/Category:The_tour_of_Doctor_Prosody,_in_search_of_the_antique_and_the_picturesque_(1821)



DOCTOR PROSODY

CORRECTING HIS PROOF IN A PRINTING OFFICE.

References (1/4)

Althaus (1973) Graphetik. In Althaus et al. (Eds.), *Lexikon der germanistischen Linguistik*. Niemeyer.

Altmann (2004) Script complexity. *Glottometrics* 8.

Browman & Goldstein (1988) Some notes on syllable structure in articulatory phonology. *Phonetica* 45.

Cutler (2015) Lexical stress in English pronunciation. In Reed & Levis (Eds.), *The handbook of English pronunciation*. John Wiley & Sons.

Eccarius & Brentari (2010) A formal analysis of phonological contrast and iconicity in sign language handshapes. *Sign Language & Linguistics* 13.2.

Eden (1961) On the formalization of handwriting. In Jakobson (Ed.), *Structure of language and its mathematical aspects*. American Mathematical Society.

Fiset et al. (2008). Features for identification of uppercase and lowercase letters. *Psychological Science* 19.11.

References (2/4)

Gibson (1969) *Principles of perceptual learning and development*. Meredith.

Gnanadesikan (2022) Amodal morphology: Applications to Brahmic scripts and Canadian Aboriginal syllabics. In Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022*, Fluxus Editions.

Kim et al. (2025) The phonology of letter shapes: Feature economy and informativeness in 43 writing systems. *Journal of Memory and Language* 14.

McCawley (1973) On the role of notation in generative phonology. Reprinted 1979 in *Adverbs, vowels and other objects of wonder*. University of Chicago Press.

Meletis & Dürscheid (2022) *Writing systems and their use: An overview of grapholinguistics*. De Gruyter Brill.

Mielke (2008) *The emergence of distinctive features*. Oxford University Press.

Miton & Morin (2021) Graphic complexity in writing systems. *Cognition* 214.

Morin (2018) Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy. *Cognitive Science* 42.2.

References (3/4)

Myers (2019) *The grammar of Chinese characters*. Routledge.

Myers (2021) Levels of structure within Chinese character constituents. In Haralambous (Ed.) *Grapholinguistics in the 21st Century: From graphemes to knowledge, Part II*. Fluxus Editions.

Nag (2014) Child and symbol factors in learning to read a visually complex writing system. *Scientific Studies of Reading* 18.5.

Peng (2017) Stroke systems in Chinese characters: A systemic functional perspective on simplified regular script. *Semiotica* 2017.218.

Peust (2006) Script complexity revisited. *Glottometrics* 12.

Primus (2004) A featural analysis of the Modern Roman Alphabet. *Written Language & Literacy* 7.2.

Primus & Wagner (2013) Buchstabenkomposition. In Ôhashi & Roussel (Eds.) *Buchstaben der Welt – Welt der Buchstaben*. Wilhelm Fink Verlag.

References (4/4)

Rezec (2009) *Zur Struktur des deutschen Schriftsystems*. Ludwig-Maximilians-Universität Ph.D. thesis.

Sablé-Meyer et al. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology* 139.

Sandler et al. (2011) The gradual emergence of phonological form in a new language. *Natural Language & Linguistic Theory* 29.2.

van Sommers (1984) *Drawing and cognition*. Cambridge University Press.

Wang (1983) *Toward a generative grammar of Chinese character structure and stroke or*der. University of Wisconsin-Madison Ph.D. thesis.

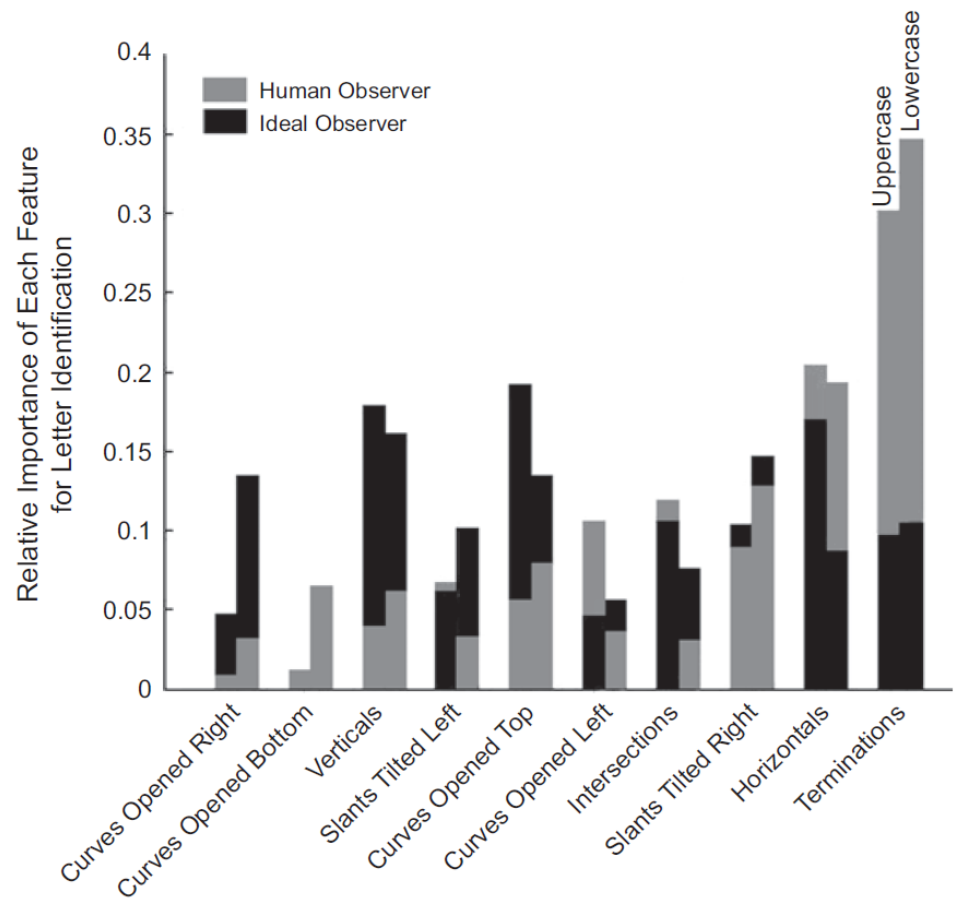
Watt (1980). What is the proper characterization of the alphabet? II: Composition. *Ars Semeiotica* 3.1.

Yang & Wang (2018) Categorical perception of Chinese characters by simplified and traditional Chinese readers. *Reading and Writing* 31.5.

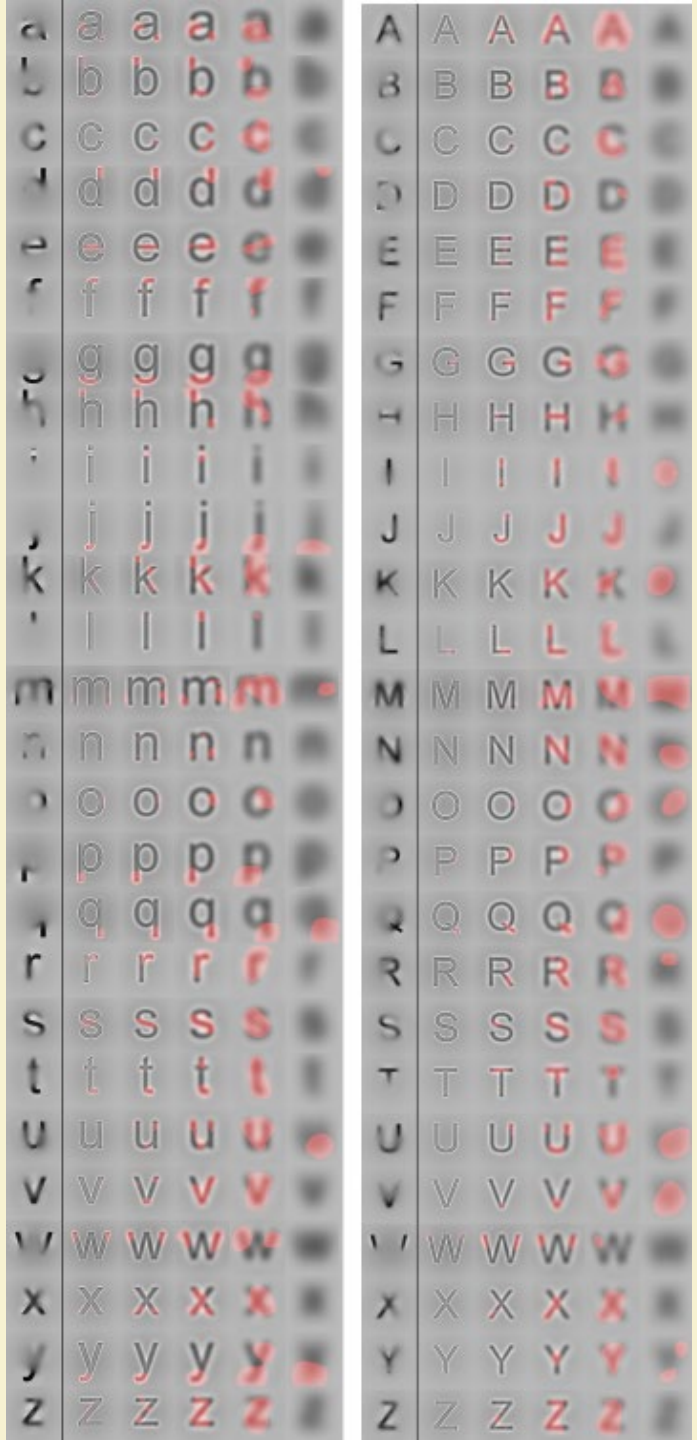
The rest of the story...

Global glyph features? (3)

- Fiset features for Latin letters



(Fiset et al. 2008)



Segmental constraints

- Stroke combinations avoid oblique angles: T X 上 *shàng* ‘over’ ᠨᠠ *ṇ-*
- But not always: A R 才 *cái* ‘talent’ ᠬᠠ *ph-*
- Non-contacting strokes seem to have more restricted axes...?
  - Usually they are parallel: 三 *sān* 川 *chuān* ‘river’ Russian: Ы *ɨ*
    (cf. hiragana: に *ni* 八 *bā* ‘eight’ 米 *mǐ* ‘rice’ - but cf. 米)
  - Or else no axis (dots): i j
  - Or system-specific default axis: Θ *theta*
    - “Falling” dots in Sinoform: 寸 *cùn* ‘inch’ katakana: シ *shi*
  - Diacritics don’t count, since they’re “affixes” (cf. Gnanadesikan 2022)
    - i vs. ì vs. í vs. ī
- So... non-parallel strokes “want” to make contact
  - Or it’s cross-“syllable” axis assimilation...? (Myers 2019, 2021)

Other open empirical questions

- Typology
  - What needs to be universal: specific elements or just general principles?
  - How should diachronic changes in glyph inventories be modeled?
- Processing
  - How does the coherence of a glyph inventory affect its learnability?
  - Are complex strokes really treated by readers and writers as sequences of simple strokes, even in scripts with extremely wiggly strokes?
- Representations (especially “prosody”)
  - Is there good justification for the syllable analog for stroke interactions?
  - What counts as “binary”? Is it just a special case of minimal contrast?