Letter shapes do not encode information as efficiently as the sounds of languages do

Yoolim Kim

Olivier Morin

Institut Jean Nicod

olivier.morin@ens.psl.eu

https://linktr.ee/oliviermorin

Helena Miton

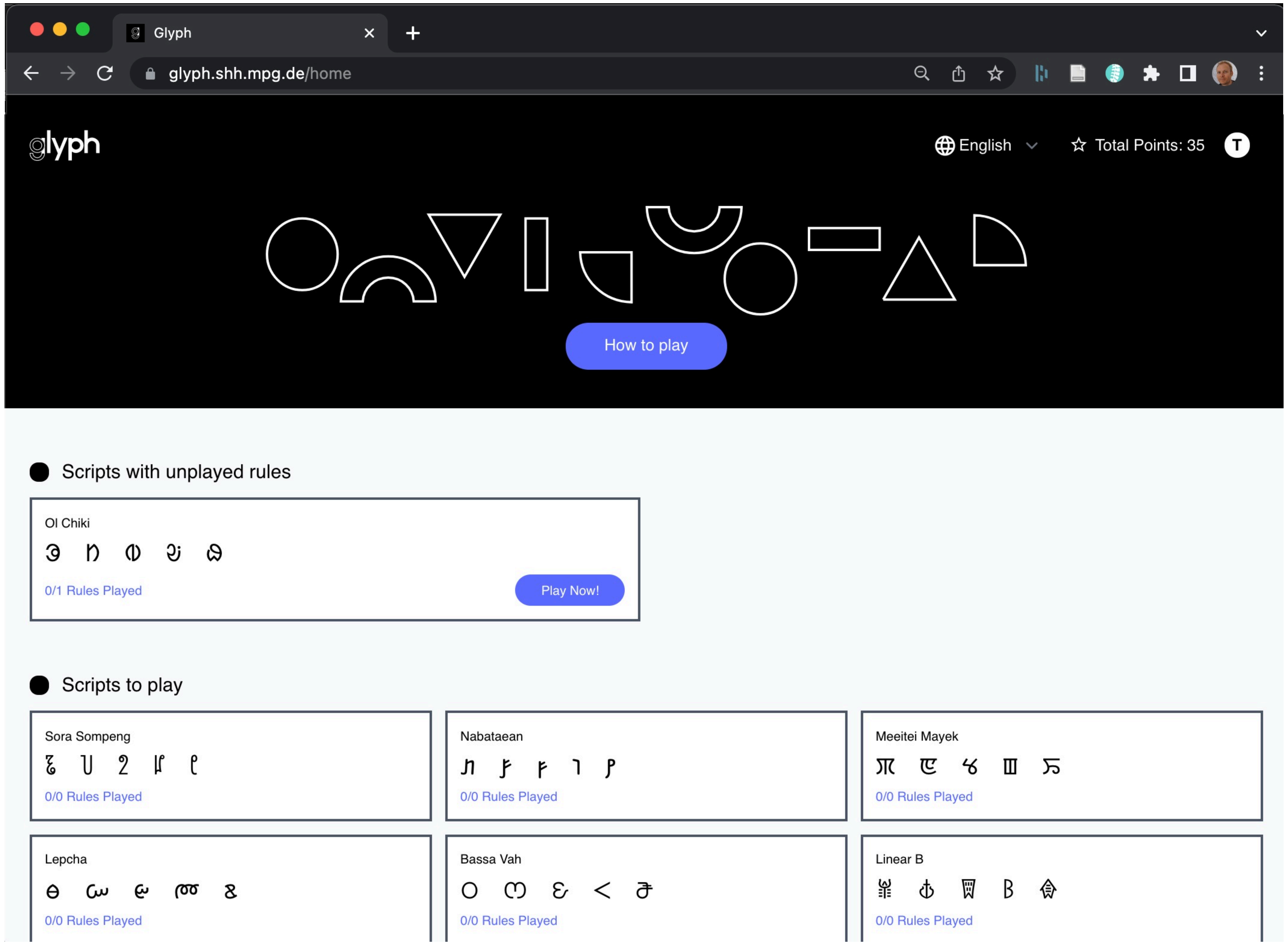
Marc Allassonnière-Tang

with

Yoolim Kim, Marc Allassonnière-Tang, Helena Miton

Thanks

to the scholars, technicians, and players behind the Glyph applet.



Scripts with unplayed rules

Ol Chiki
0/1 Rules Played

Scripts to play

Sora Sompeng
0/0 Rules Played

Nabataean
0/0 Rules Played

Meeitei Mayek
0/0 Rules Played

Lepcha
0/0 Rules Played

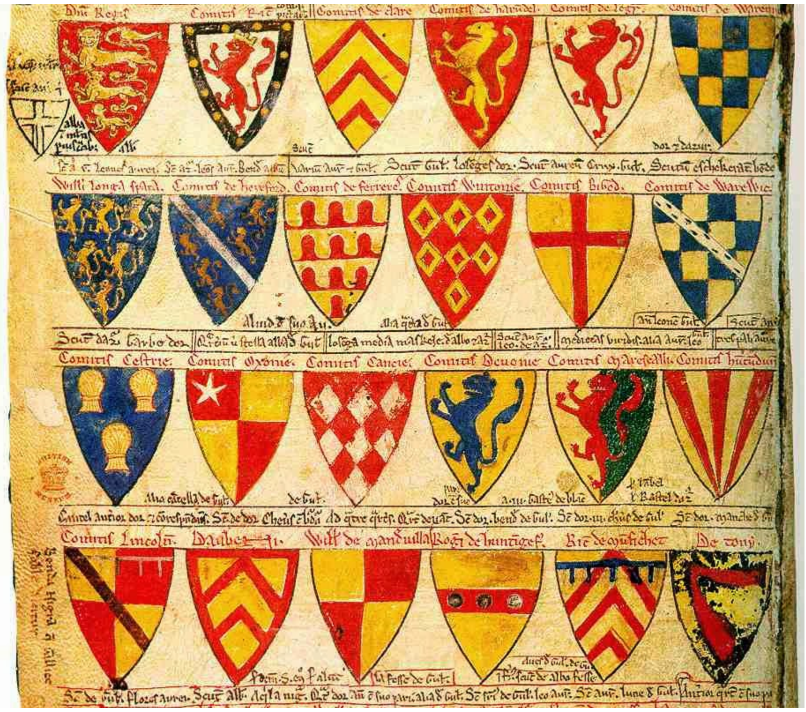
Bassa Vah
0/0 Rules Played

Linear B
0/0 Rules Played

a small caveat

How do images encode information?

Pavlek et al. 2019

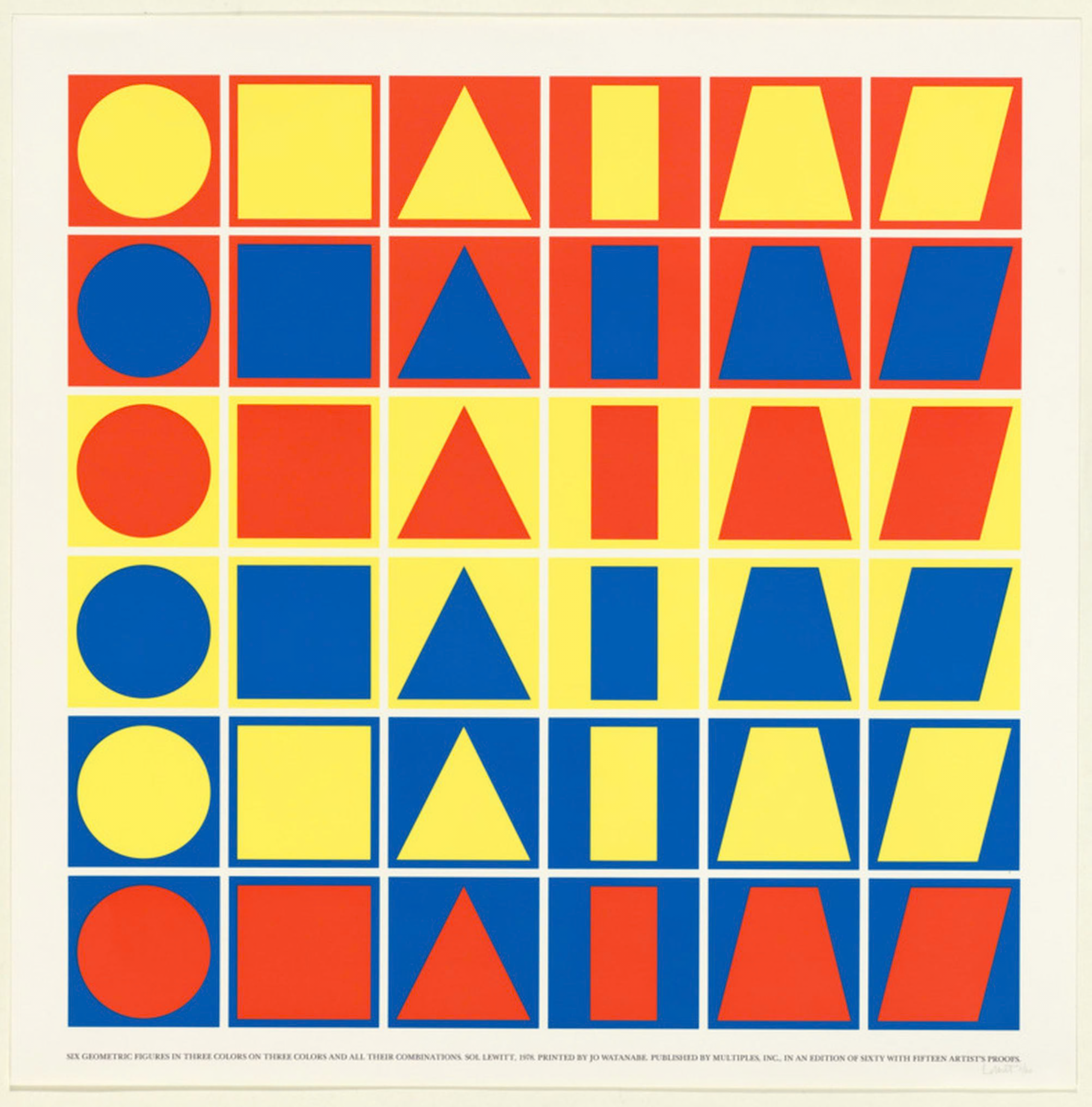


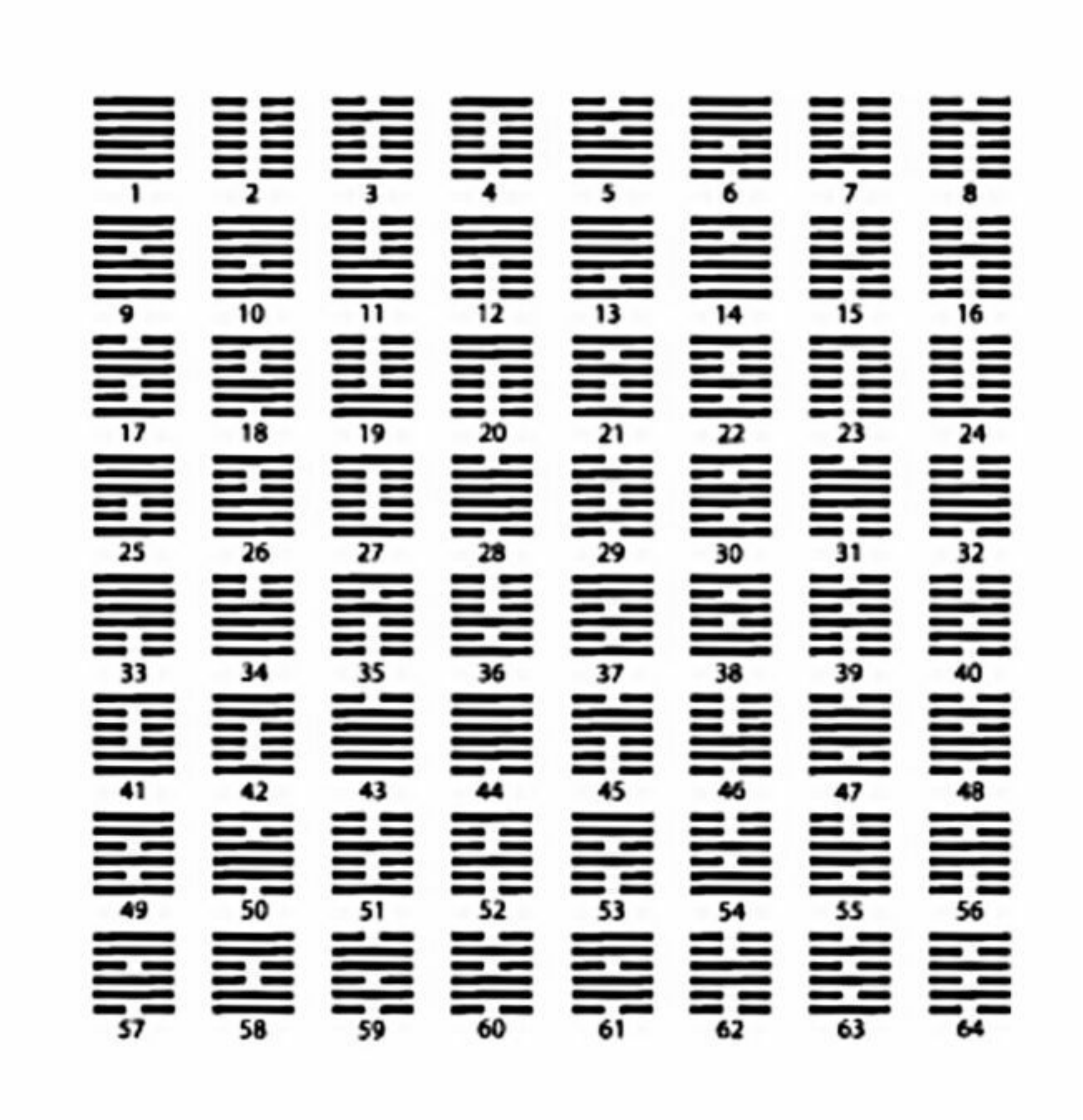
Morin & Miton 2018

Youngblood Miton & Morin 2023

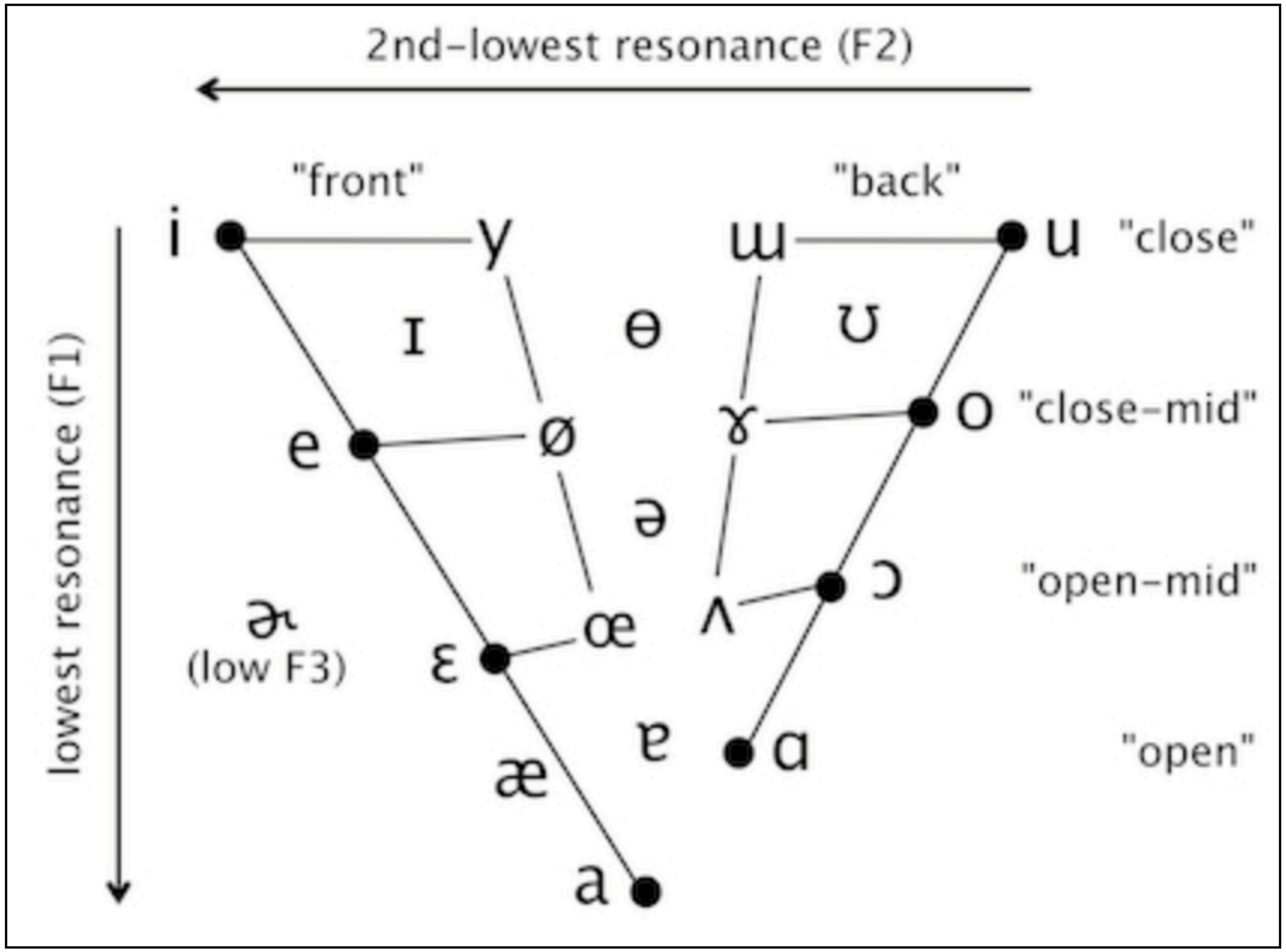
Combinatoriality:

using a few dimensions of variation to create many diverse forms

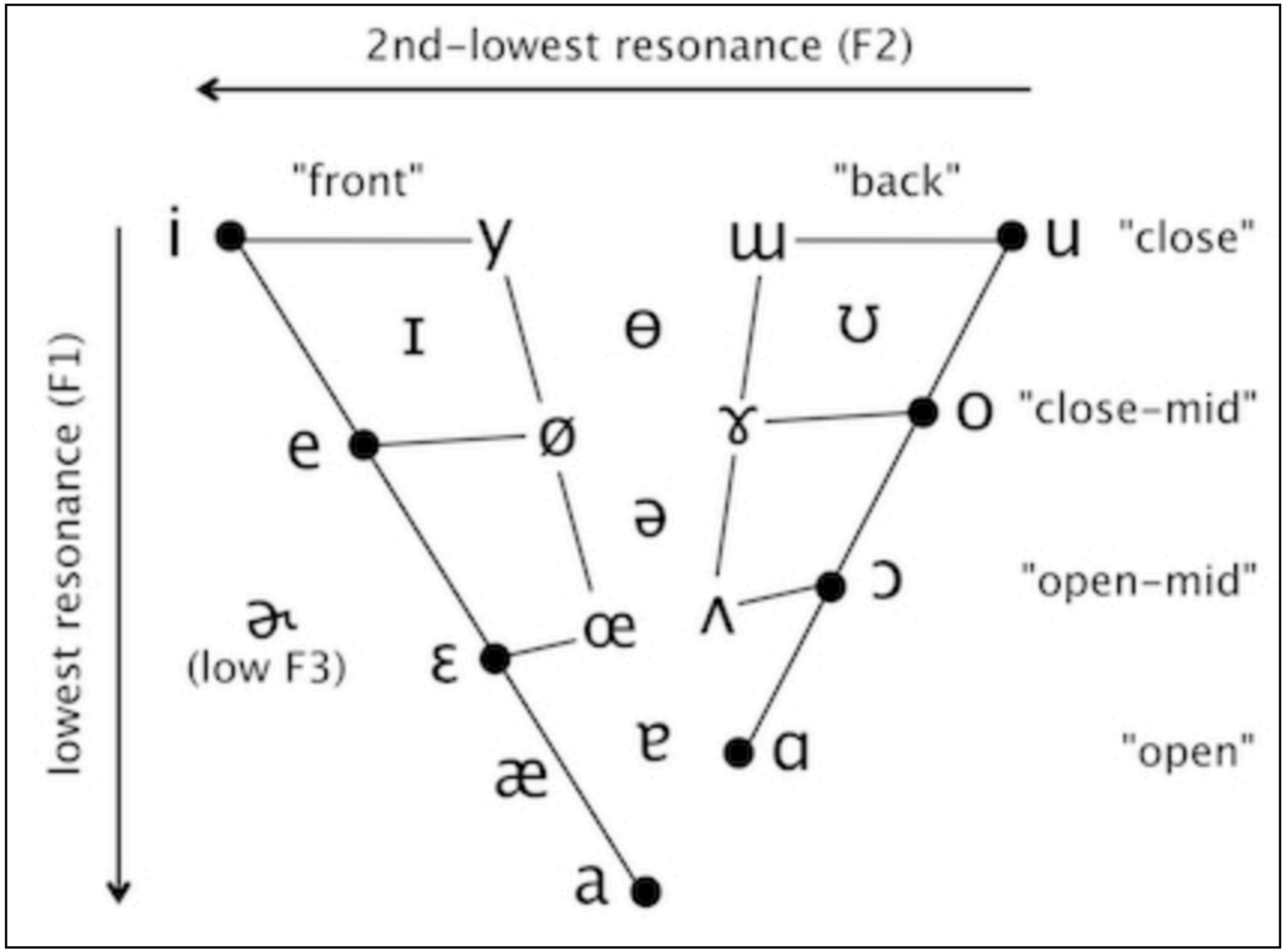

SIX GEOMETRIC FIGURES IN THREE COLORS ON THREE COLORS AND ALL THEIR COMBINATIONS. SOL LEWITT, 1978. PRINTED BY JO WATANABE. PUBLISHED BY MULTIPLES, INC., IN AN EDITION OF SIXTY WITH FIFTEEN ARTIST'S PROOFS.



Speech sounds are combinatorial



Speech sounds are combinatorial



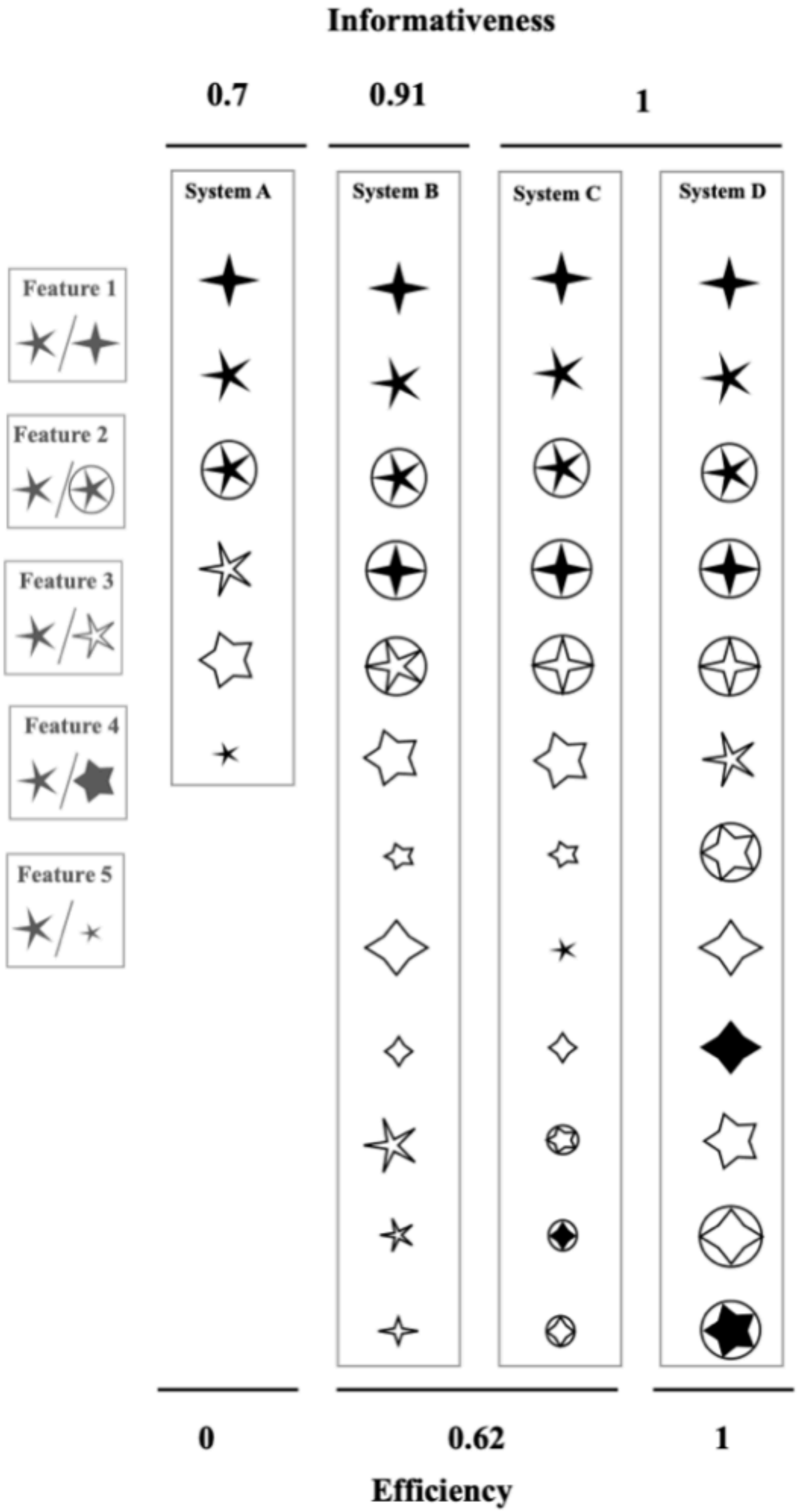
=> easier to learn

=> easier to encode

=> easier to pronounce

(Martinet 1971)

Two dimensions of combinatoriality

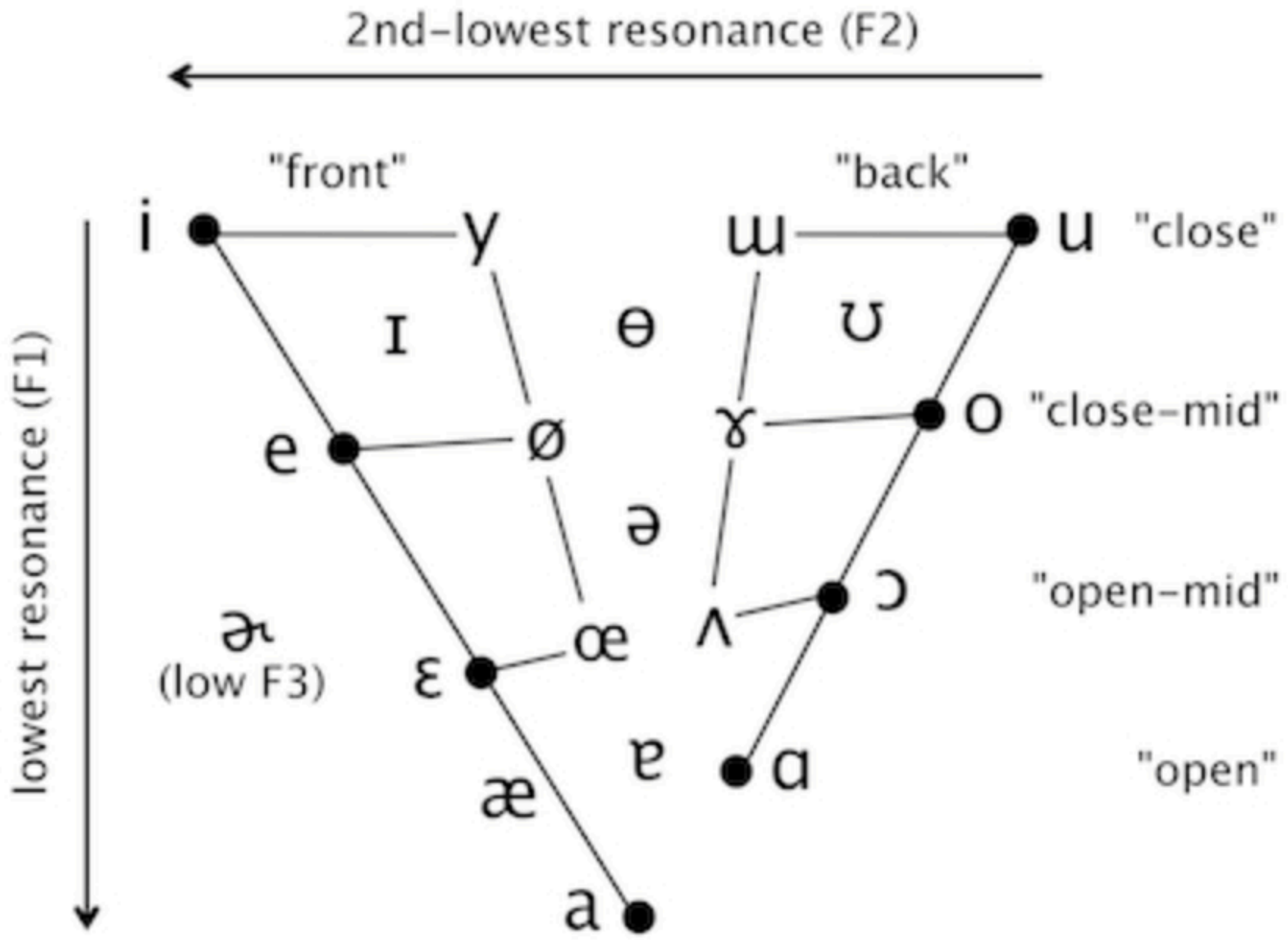


Feature informativeness: What are the odds that two random symbols differ on a given feature? For binary features, maximally high if 50% of symbols have one value and the other symbols have the other value for this feature.

Feature efficiency: What is the smallest number of features needed to describe all the symbols in a system? And how many symbols does the system generate from these features?

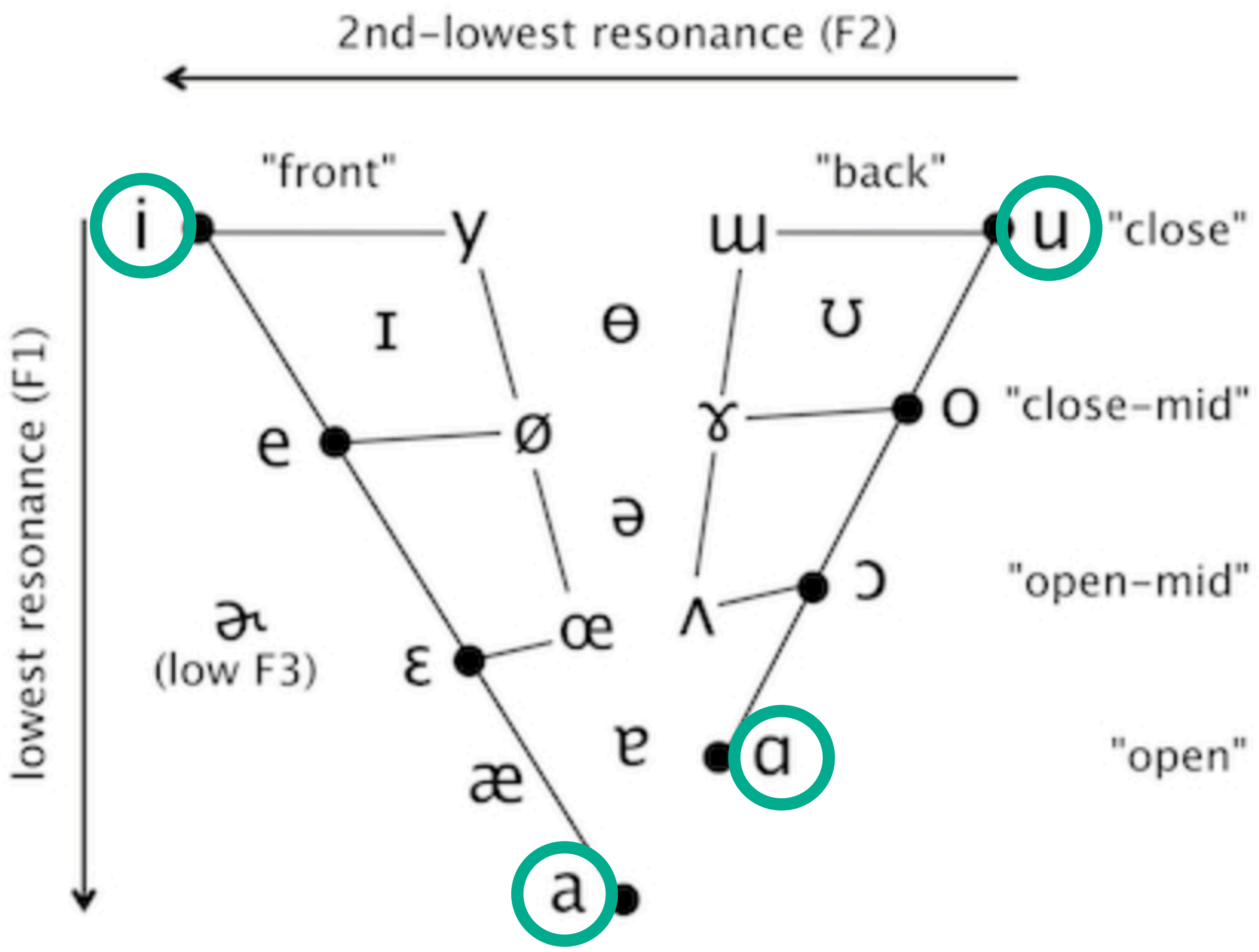
Phonological features are **efficient** and informative.

A few features suffice to describe many different phonemes, within and across languages ("feature economy"—Clements 2003).



The vowel space in three features
(from Wikipedia, not including nasal vowels)

Phonological features are efficient and **informative**.



A typical vowel system:
Symmetrical (Dupoux & Dunbar 2016), with optimal dispersal (Lindblom 1986).

Each feature discriminates as many sounds as possible (50%): the features are maximally informative.

What about writing systems?

Letters in writing systems play a very similar role to phonemes in spoken language, and obey similar constraints (they need to be easily learnt, stored, processed).

And some writing systems seem clearly combinatorial...

... but relevant empirical work is scarce.

Our prediction:

Writing is not as combinatorial as speech,
because movements of the hand
are less constrained than movements of the tongue.

(Sandler 2008: a similar point for sign language.)

Glyph: Crowdsourcing a typology of letter shapes

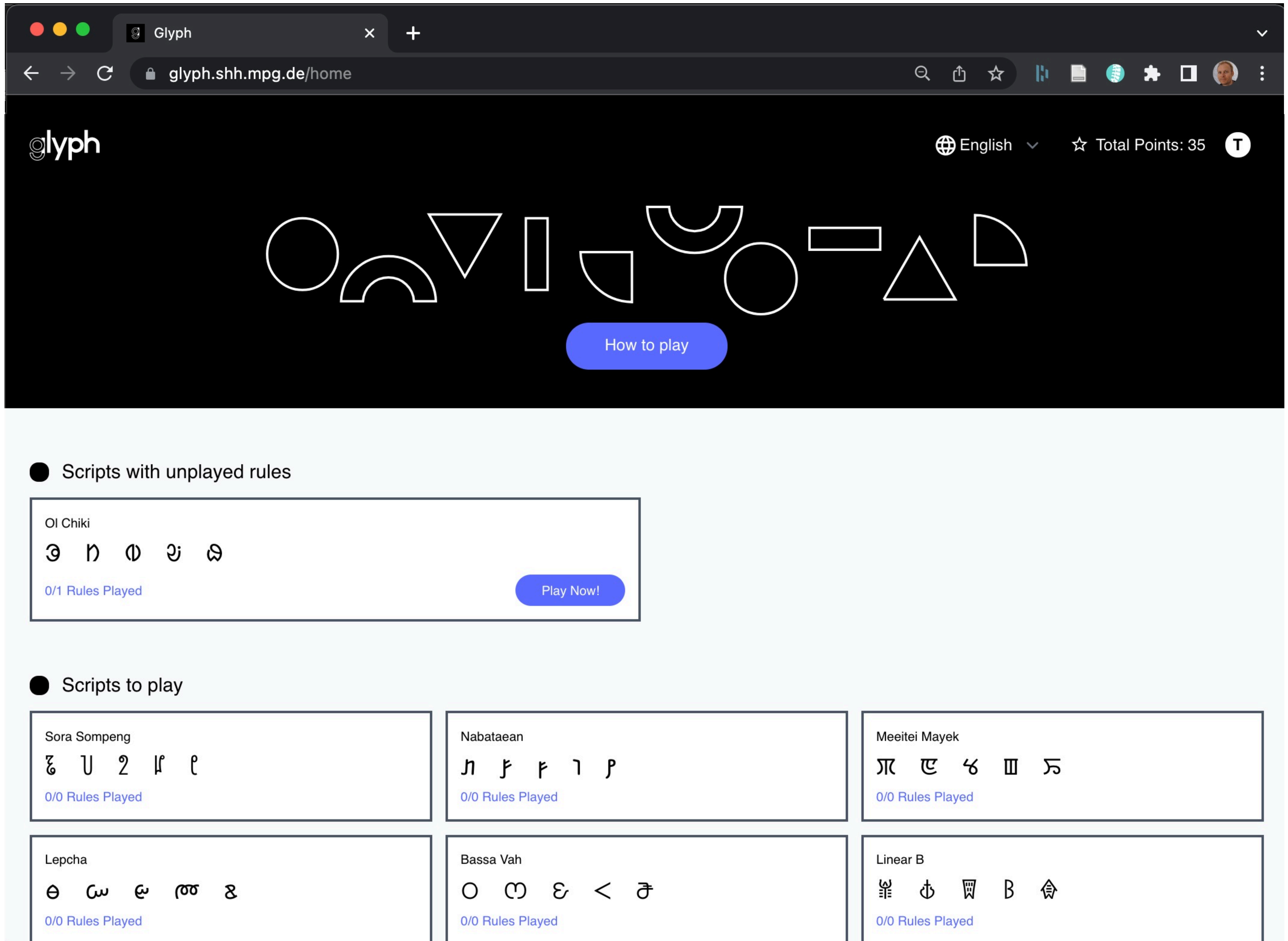
glyph.shh.mpg.de

Linguists have the IPA, but our field lacks a *systematic* classification of letter shapes, allowing for rigorous comparative and diachronic studies.

Glyph, a citizen science project, aims to build this.

> 4,500 participants

~ 70,000 classifications proposed



Glyph ×

glyph.shh.mpg.de/home

glyph

English

☆ Total Points: 35

How to play

● Scripts with unplayed rules

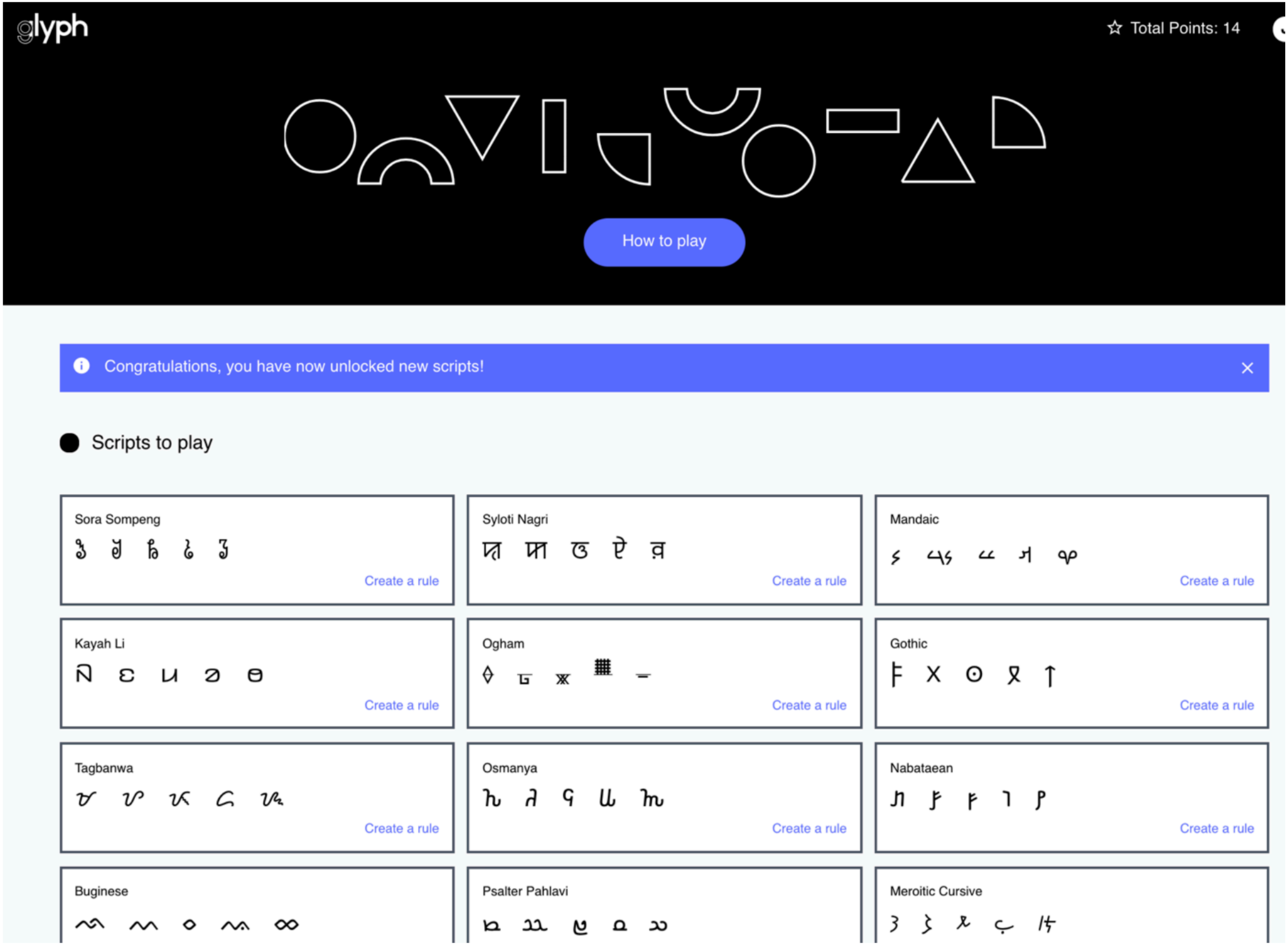
Ol Chiki

0/1 Rules Played

Play Now!

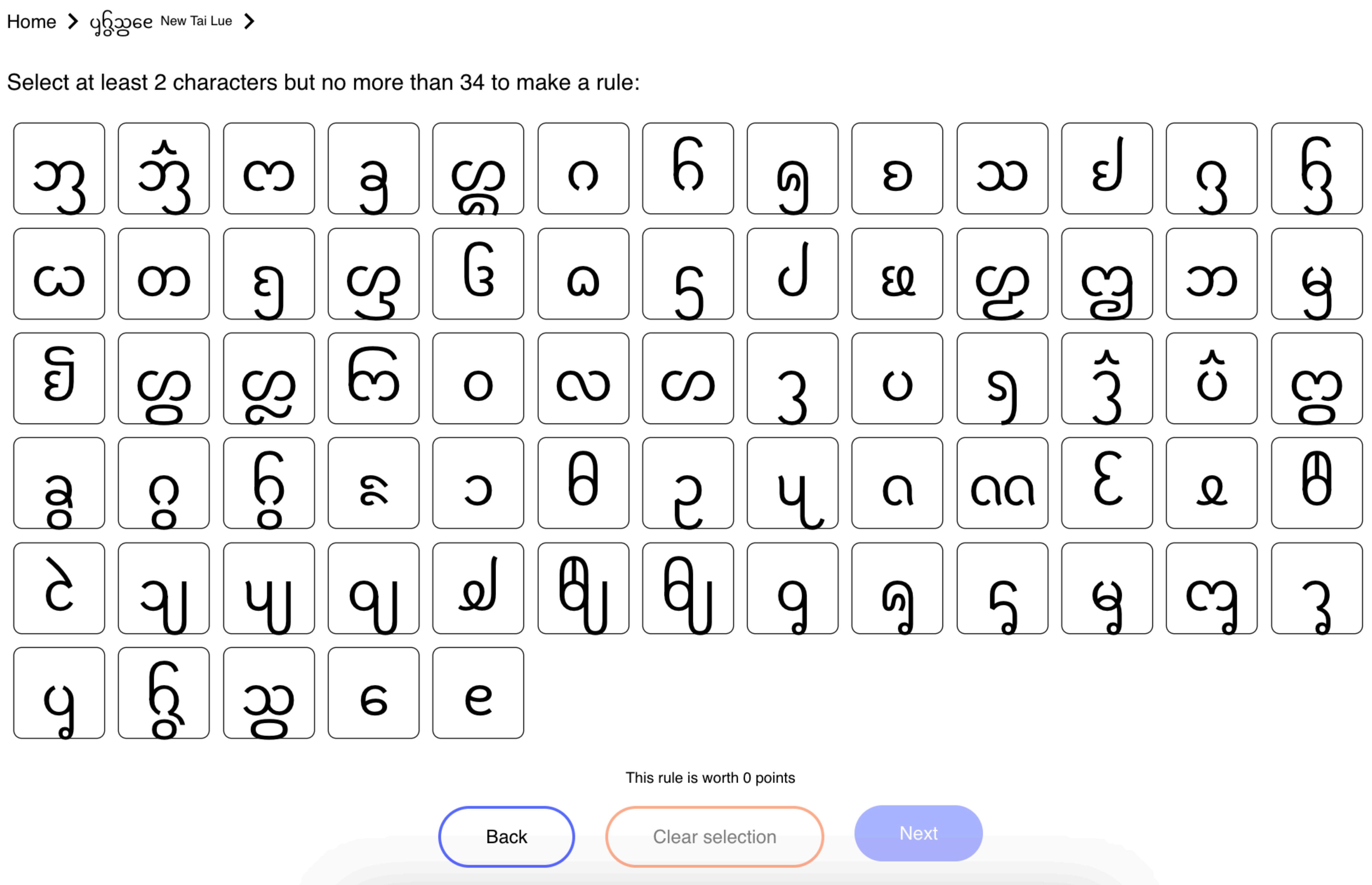
● Scripts to play

Sora Sompeng 0/0 Rules Played	Nabataean 0/0 Rules Played	Meeitei Mayek 0/0 Rules Played
Lepcha 0/0 Rules Played	Bassa Vah 0/0 Rules Played	Linear B 0/0 Rules Played



Congratulations, you have now unlocked new scripts!

● Scripts to play



Home > New Tai Lue >

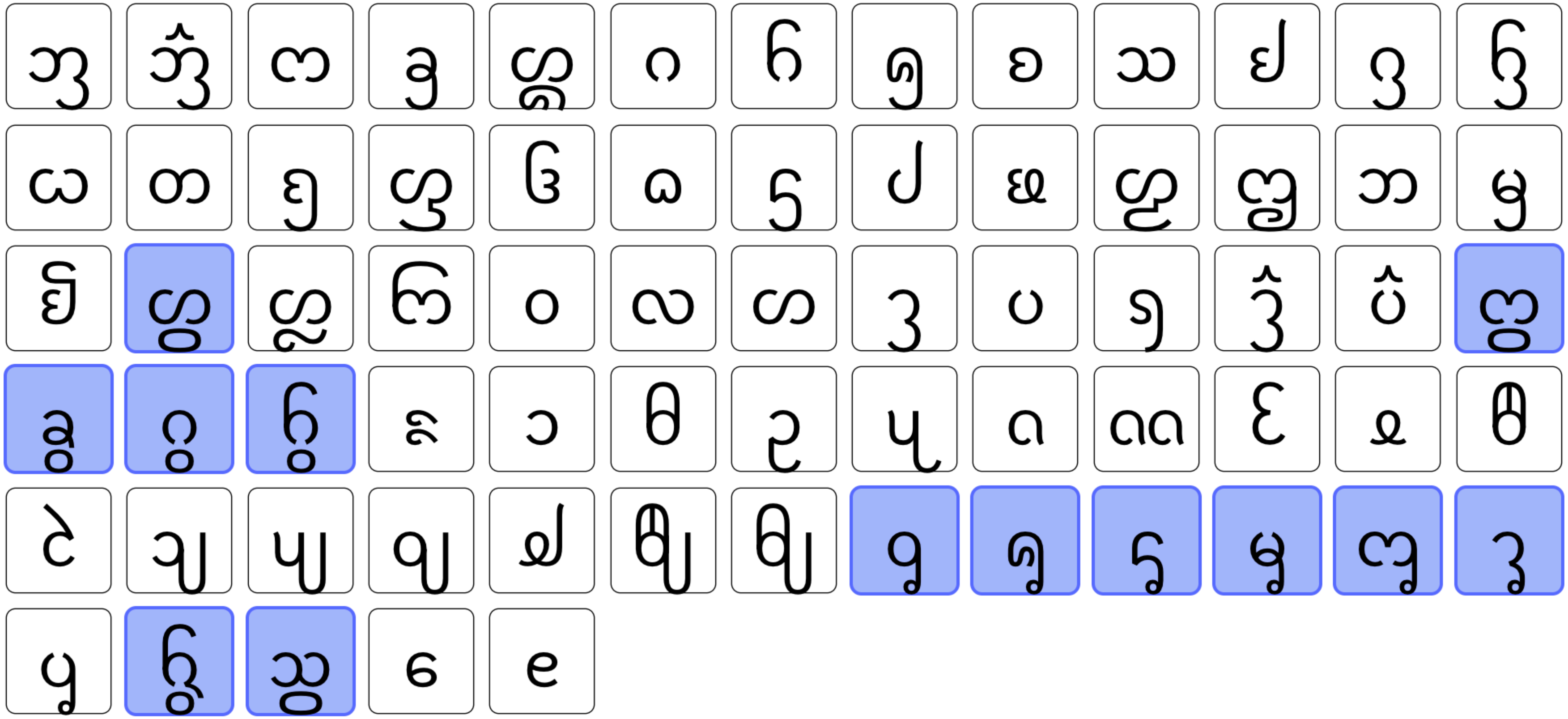
Select at least 2 characters but no more than 34 to make a rule:

This rule is worth 0 points

New Tai Lue script for Tai Lü (Yunnan, China)

Home > New Tai Lue >

Select at least 2 characters but no more than 34 to make a rule:

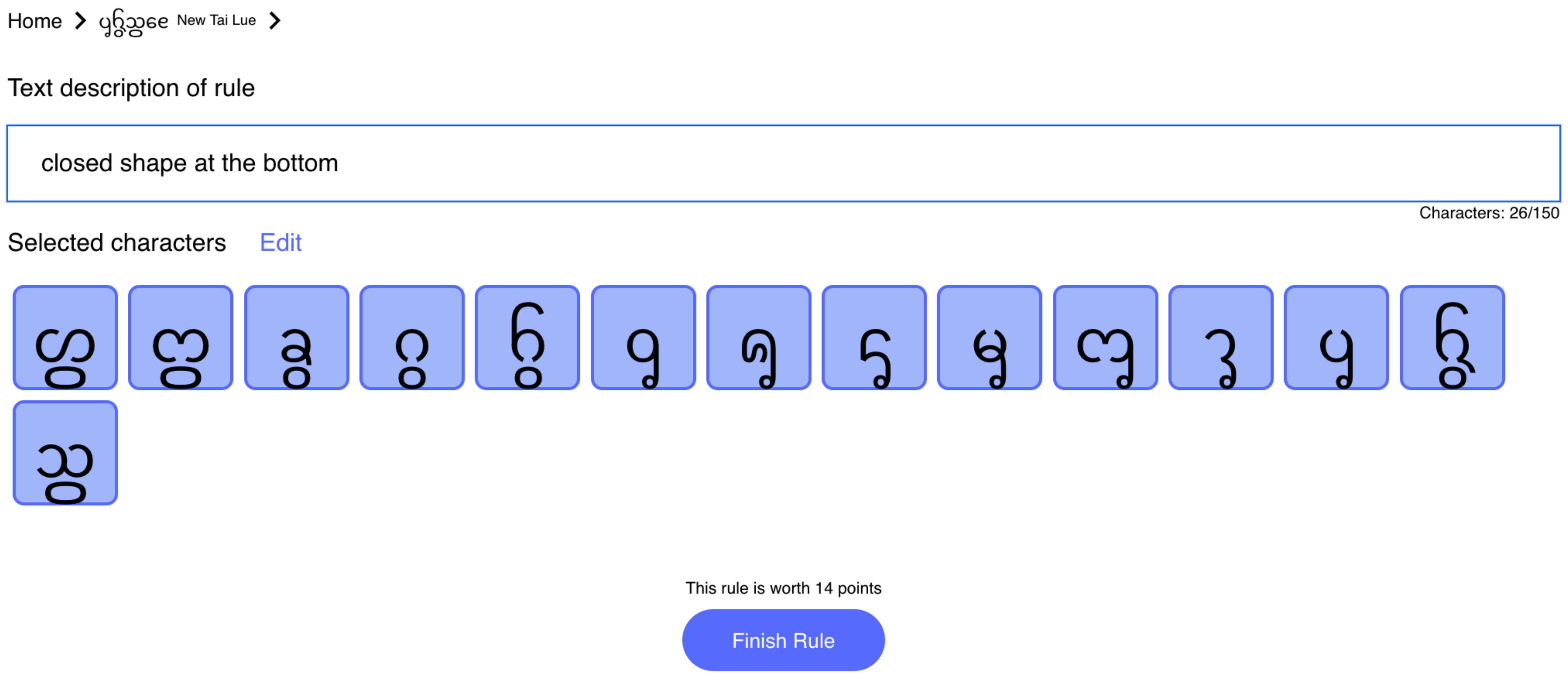


This rule is worth 13 points

Back

Clear selection

Next



Home > ᦟᦲᧅᦑᦺᦟᦹᧉ New Tai Lue >

Text description of rule

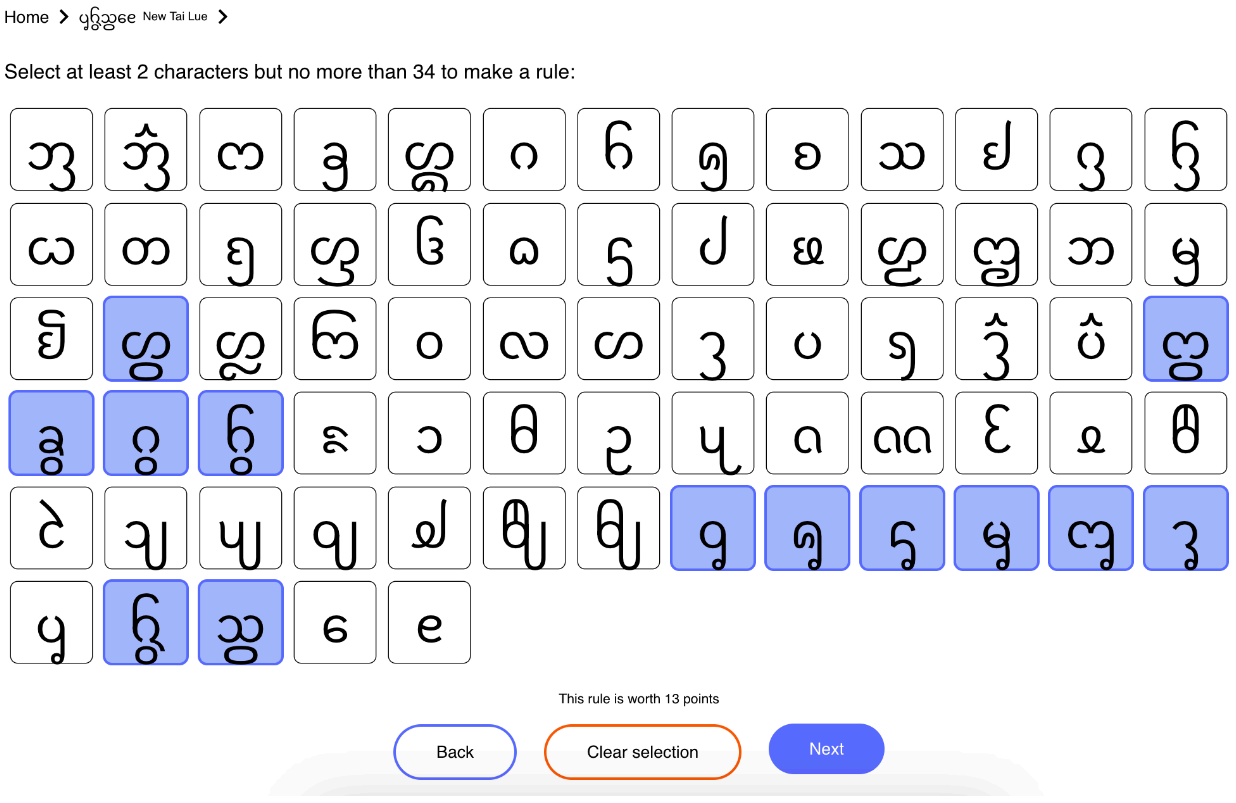
closed shape at the bottom

Characters: 26/150

Selected characters

This rule is worth 14 points

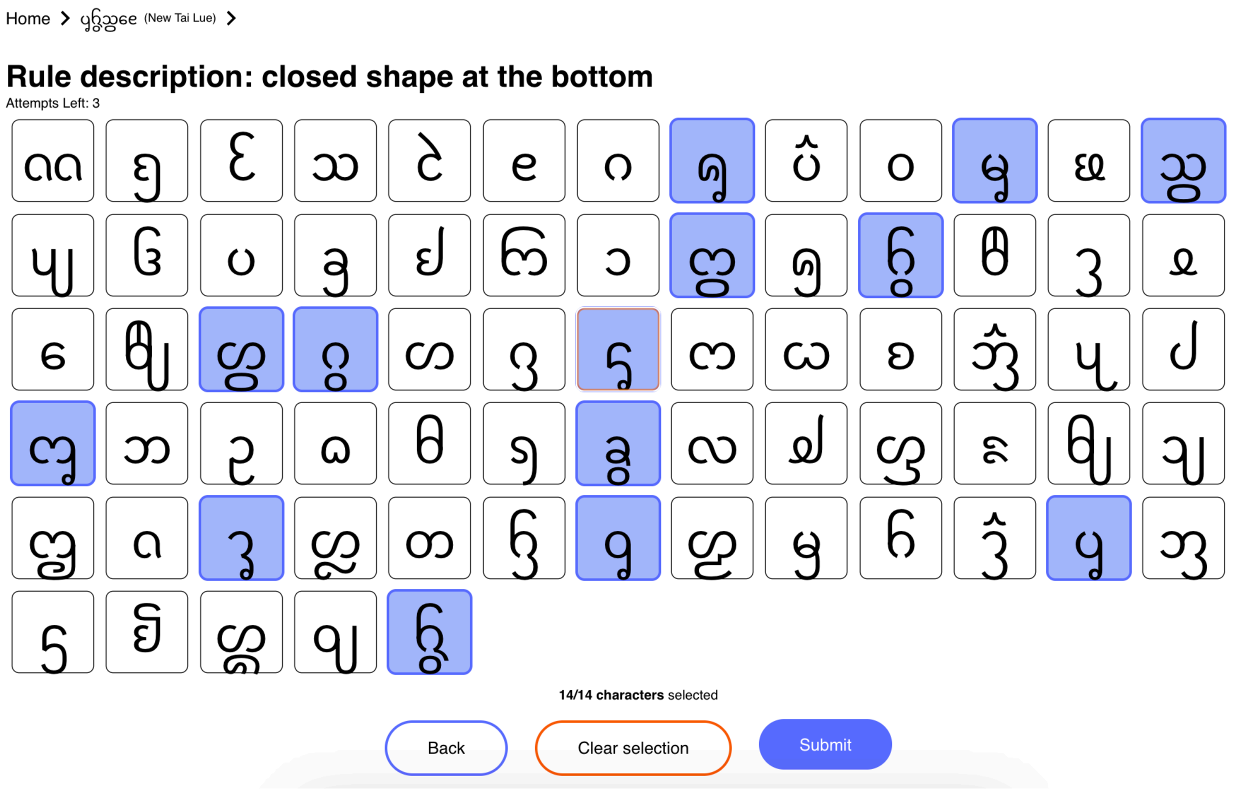
3 minutes



Home > New Tai Lue >

Select at least 2 characters but no more than 34 to make a rule:

This rule is worth 13 points



Home > (New Tai Lue) >

Rule description: closed shape at the bottom

Attempts Left: 3

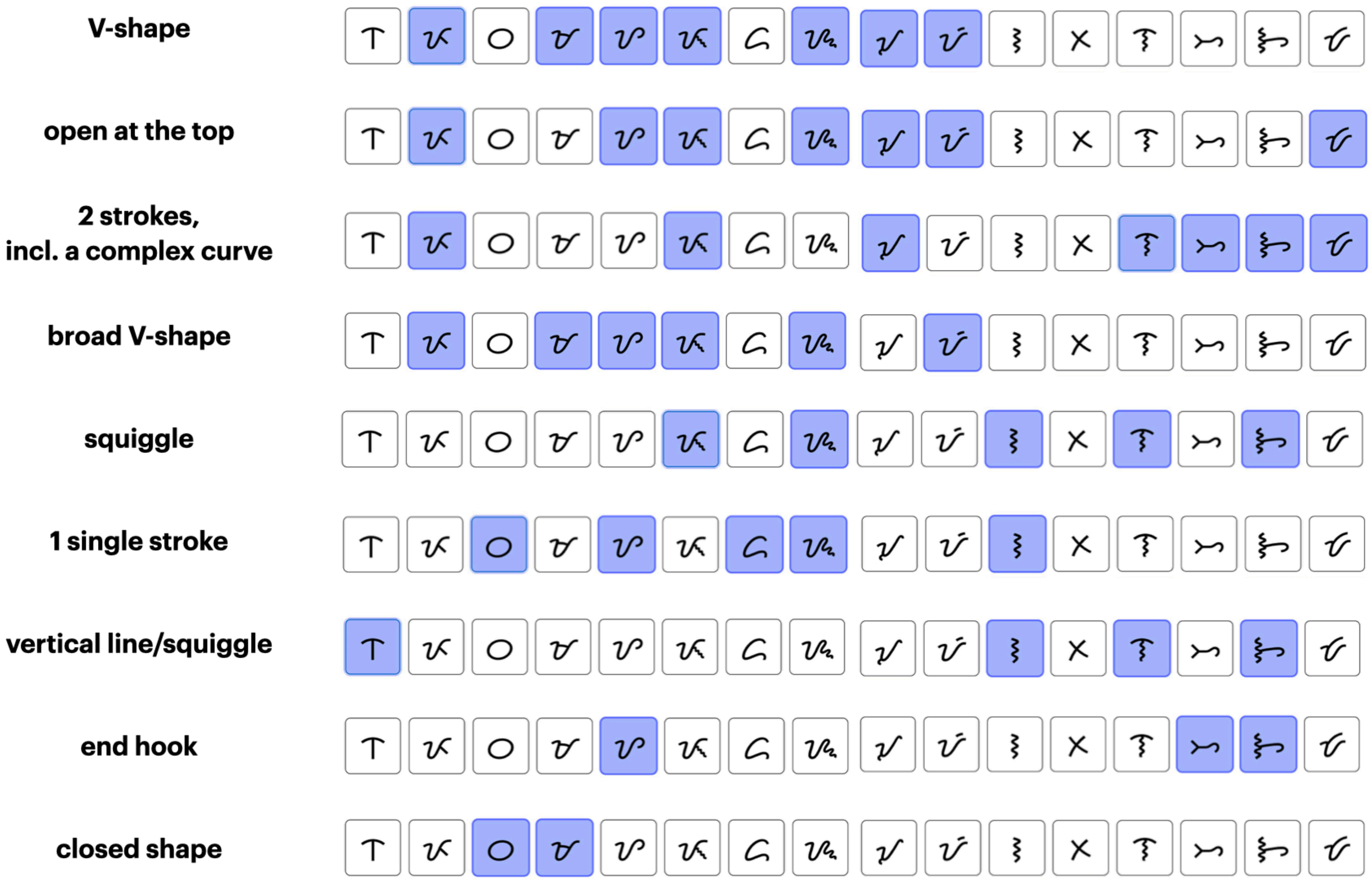
14/14 characters selected

You passed!

You earned 14 points!

Continue

Selecting the best set of classifications for each script



V-shape

open at the top

2 strokes,
incl. a complex curve

broad V-shape

squiggle

1 single stroke

vertical line/squiggle

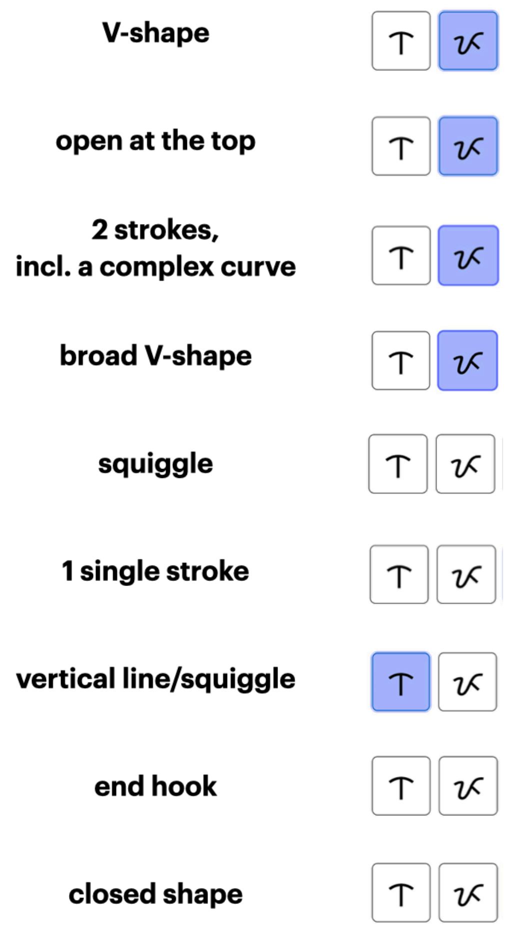
end hook

closed shape

Tagbanwa (Philippino script)

A decision tree classification algorithm selects, among all classifications produced by Glyph users, the smallest number of classifications capable of identifying the largest number of letters.

Glyph classifications are meaningful: They predict people's judgements for similarity between letters



V-shape

open at the top

2 strokes,
incl. a complex curve

broad V-shape

squiggle

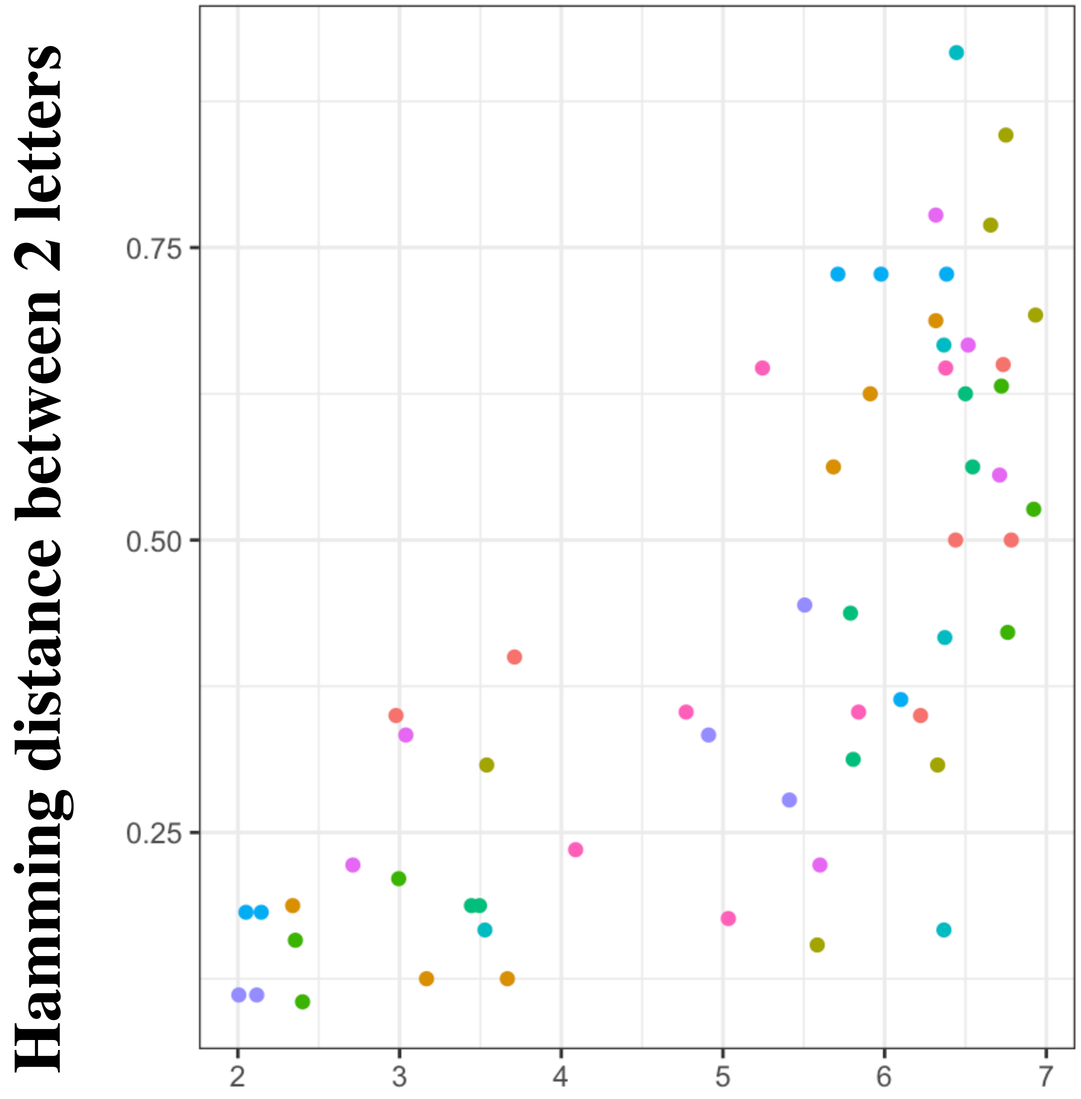
1 single stroke

vertical line/squiggle

end hook

closed shape

Hamming distance between ⊤ and ϰ : 5.
Normalised Hamming distance: 5/9 = 0.555...



Perceptual dissimilarity score
(average of 180 participants)

One point = one pair of letters

Pearson's r = 0.75

Feature economy: Mackie & Mielke's 'relative efficiency'

Compares the number of features used by a system to the minimal and the maximal number of features needed to distinguish all symbols.

$$(1)\ RE = \sqrt{\frac{F - F_{min}}{F_{max} - F_{min}}}$$

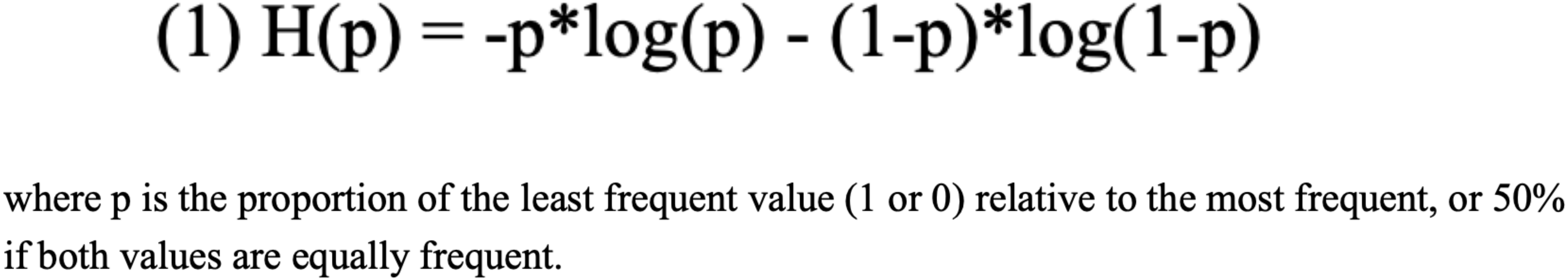
$$(2)\ F_{min} = \lceil log_2(S) \rceil$$

$$(3)\ F_{max} = S - 1$$

Clement's feature economy (F/S) is simpler but yields artifacts (e.g. two perfectly economical inventories differing in size will vary in Clement's FE). FE is robust to variation in the absolute magnitude of S.

Feature informativeness

We compute the informativeness of each feature using Shannon entropy:

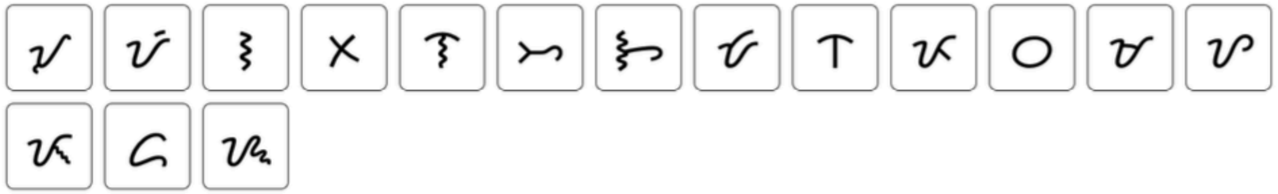


$$(1)\ H(p) = -p*\log(p) - (1-p)*\log(1-p)$$

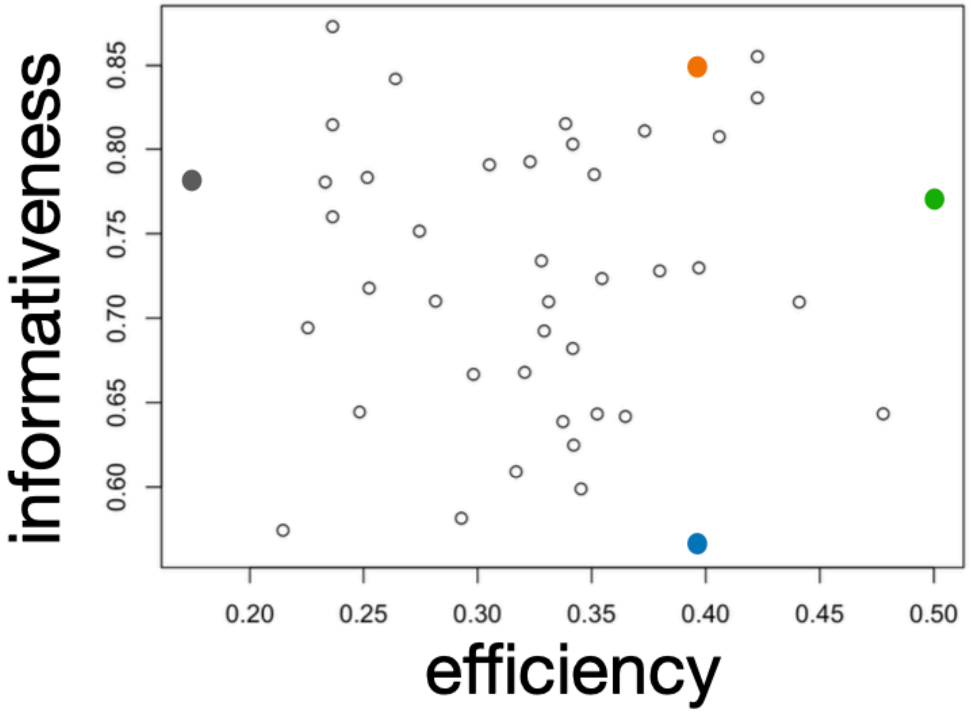
where p is the proportion of the least frequent value (1 or 0) relative to the most frequent, or 50% if both values are equally frequent.

then average this over all features in a script / language.

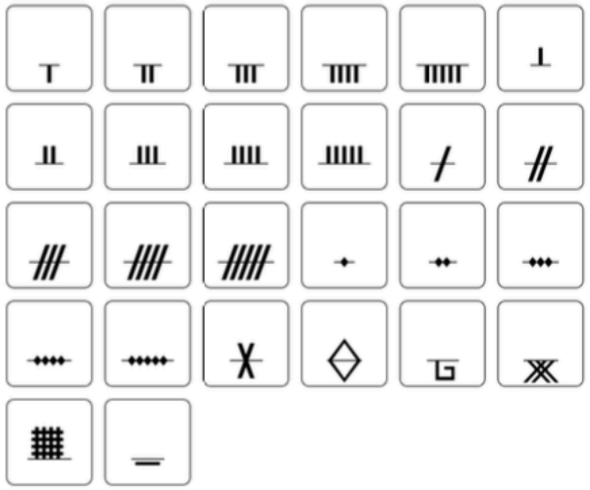
Tagbanwa



Kayah Li



Ogham



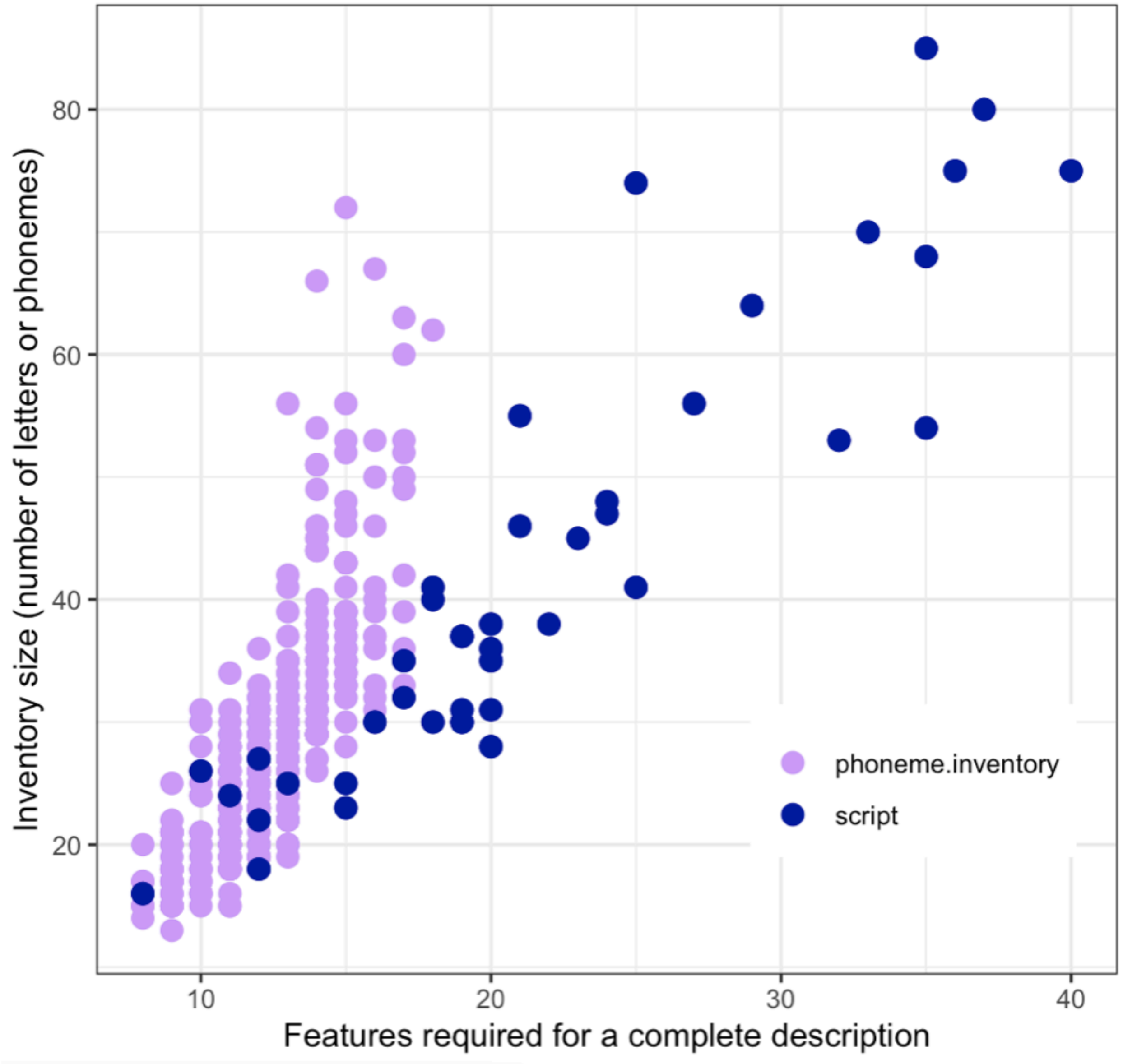
Cherokee

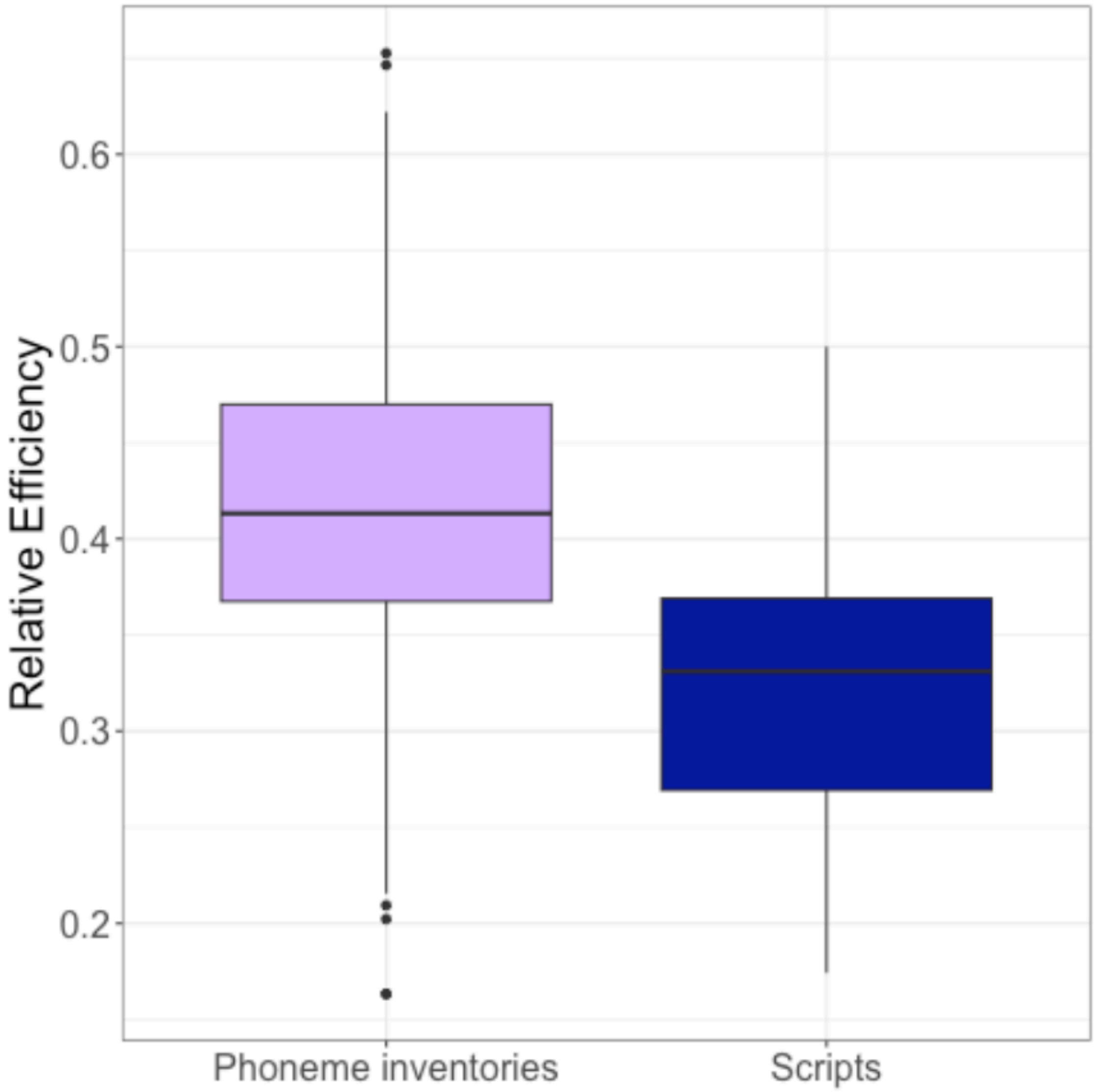
Datasets:

43 Glyph scripts (chosen to be a representative sample of the typological, areal & semiotic diversity of scripts).

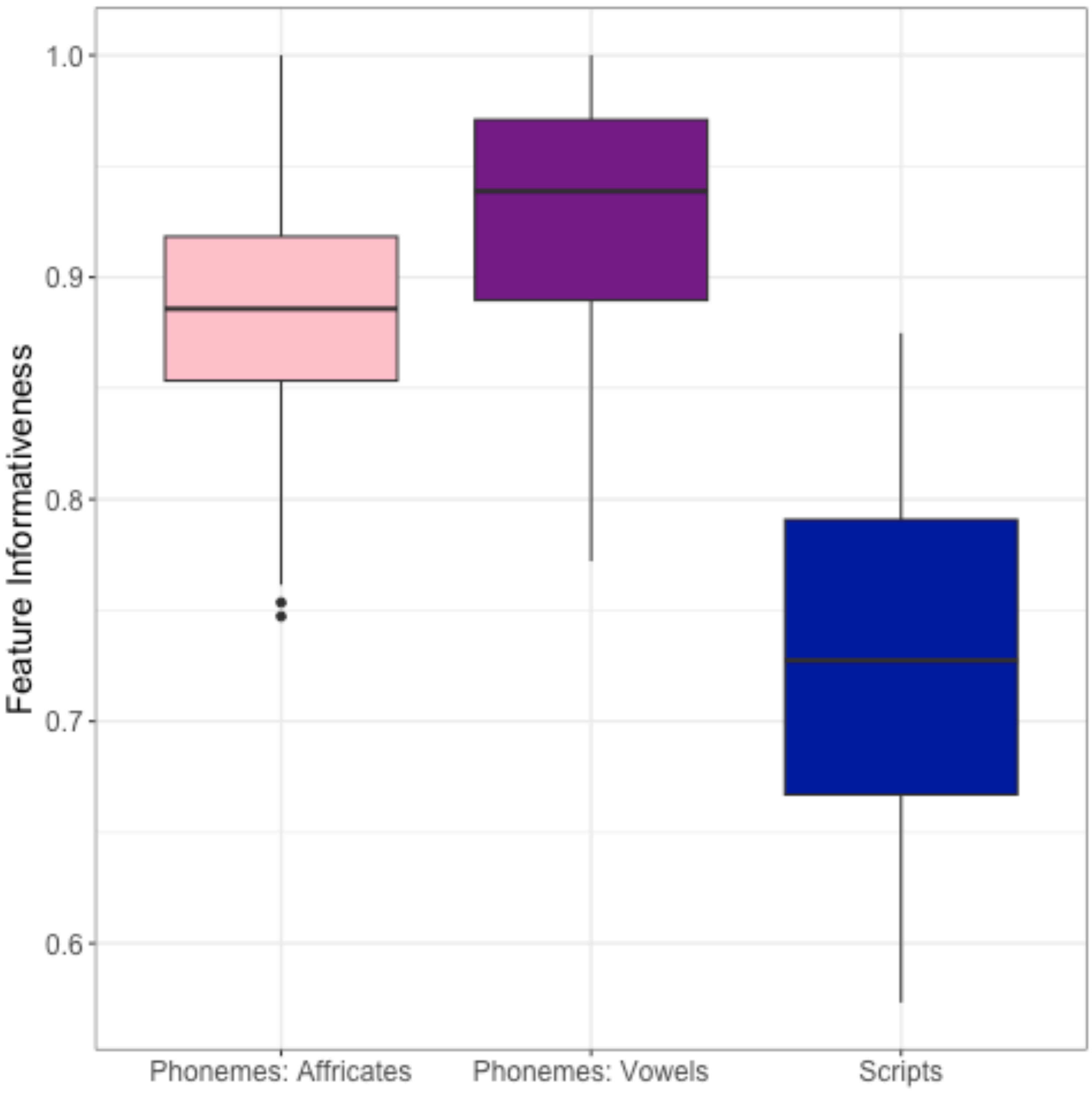
500+ phoneme inventories from Mielke's P-base dataset, each phoneme being coded as 24 binary phonemic features.

Prediction 1: Feature economy is lower in scripts





Prediction 2: Feature informativeness is lower in scripts



Null model: Predicting distinctiveness with family

Test model: Predicting distinctiveness with family + type (script or language)

ΔAIC = 104

(Also works when controlling for number of letters/phonemes)

Letter shapes are only weakly combinatorial, compared to sounds.

Spoken languages have many unique distinctive sounds because they combine a few phonological features very informatively.

Written languages have many distinctive letters because they combine many visual features, less informatively.

Thanks!

https://linktr.ee/oliviermorin

olivier.morin@ens.psl.eu

slides↑

Generalist features

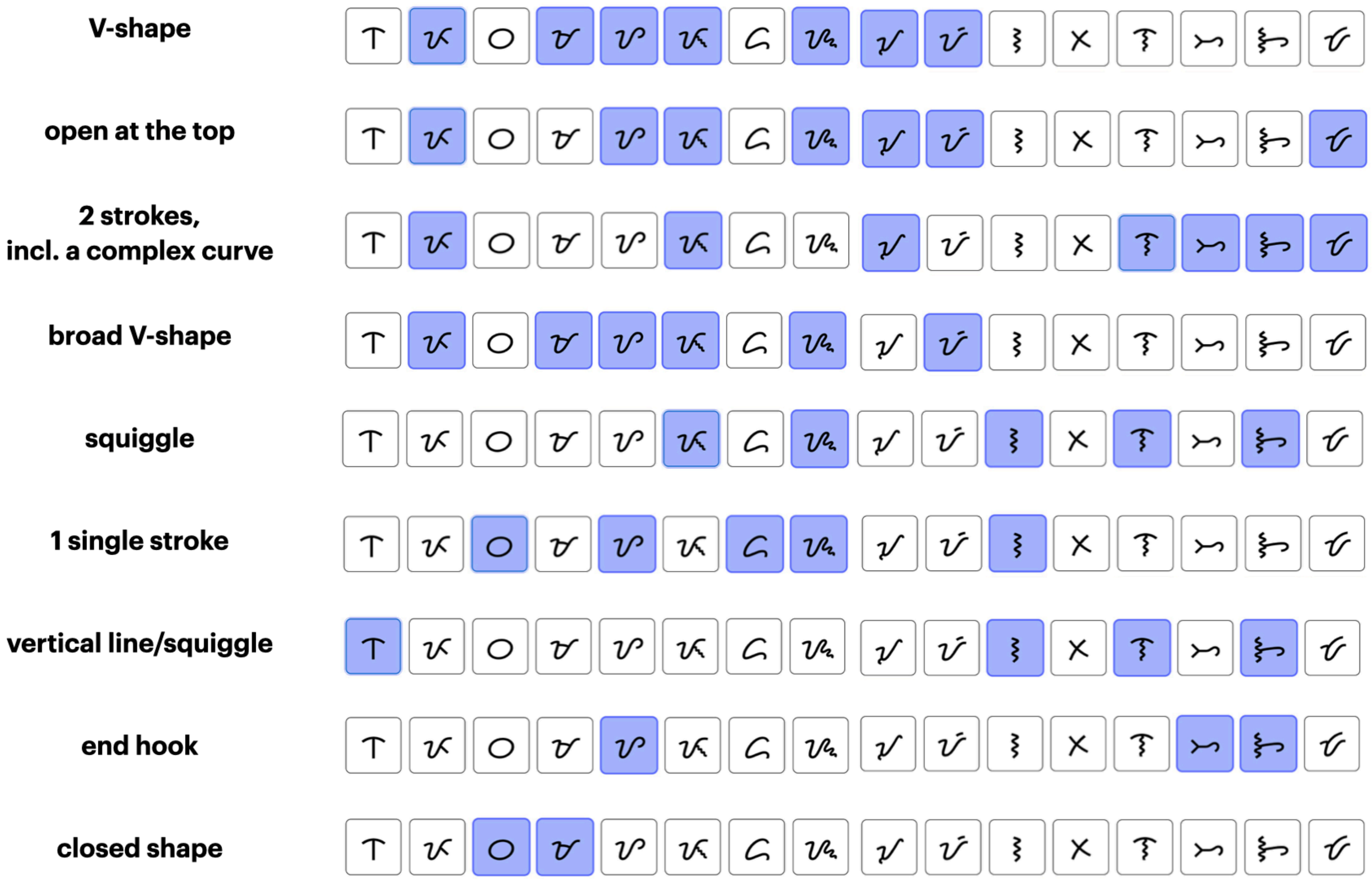
Feature economy and informativeness in phonemes: the nitty-gritty details

To keep the comparison between scripts and phonemes equal, we applied our Best Set algorithm to both phoneme inventories and scripts: we considered only the features necessary to give a complete description of the script / inventory (= a unique combination of feature values for each letter / phoneme).

When computing feature informativeness, we only want to consider the features that can logically take a contrastive value for a group of phonemes, not those that take the 0 value because they don't apply. For instance, the feature "sibilant" / "non sibilant"does not apply to vowels.

=> Feature informativeness only measured for sub-inventories: vowels, and affricate consonants, separately.

So far we used bespoke classifications for each script (the "best set" of features).



V-shape

open at the top

2 strokes,
incl. a complex curve

broad V-shape

squiggle

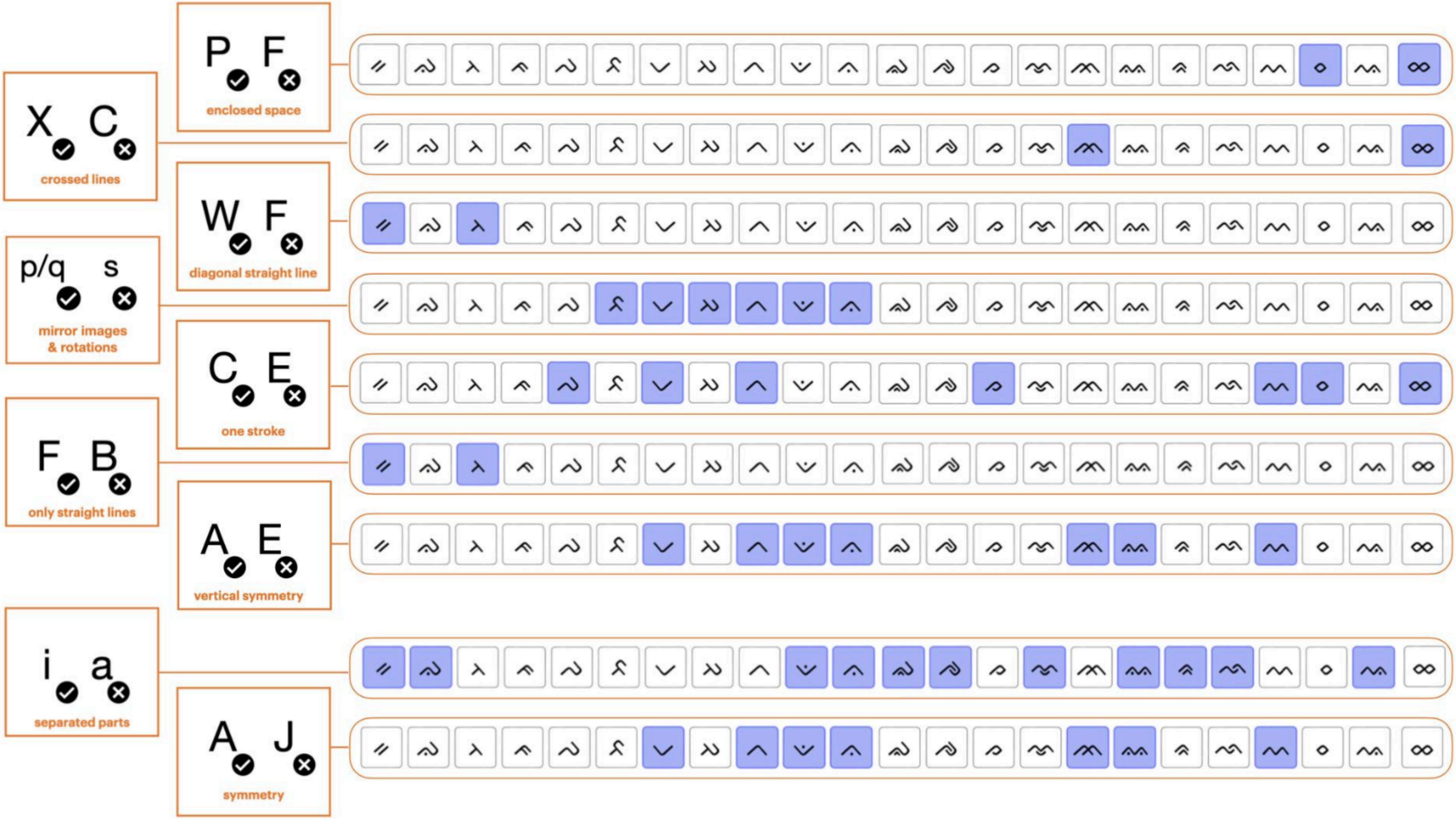
1 single stroke

vertical line/squiggle

end hook

closed shape

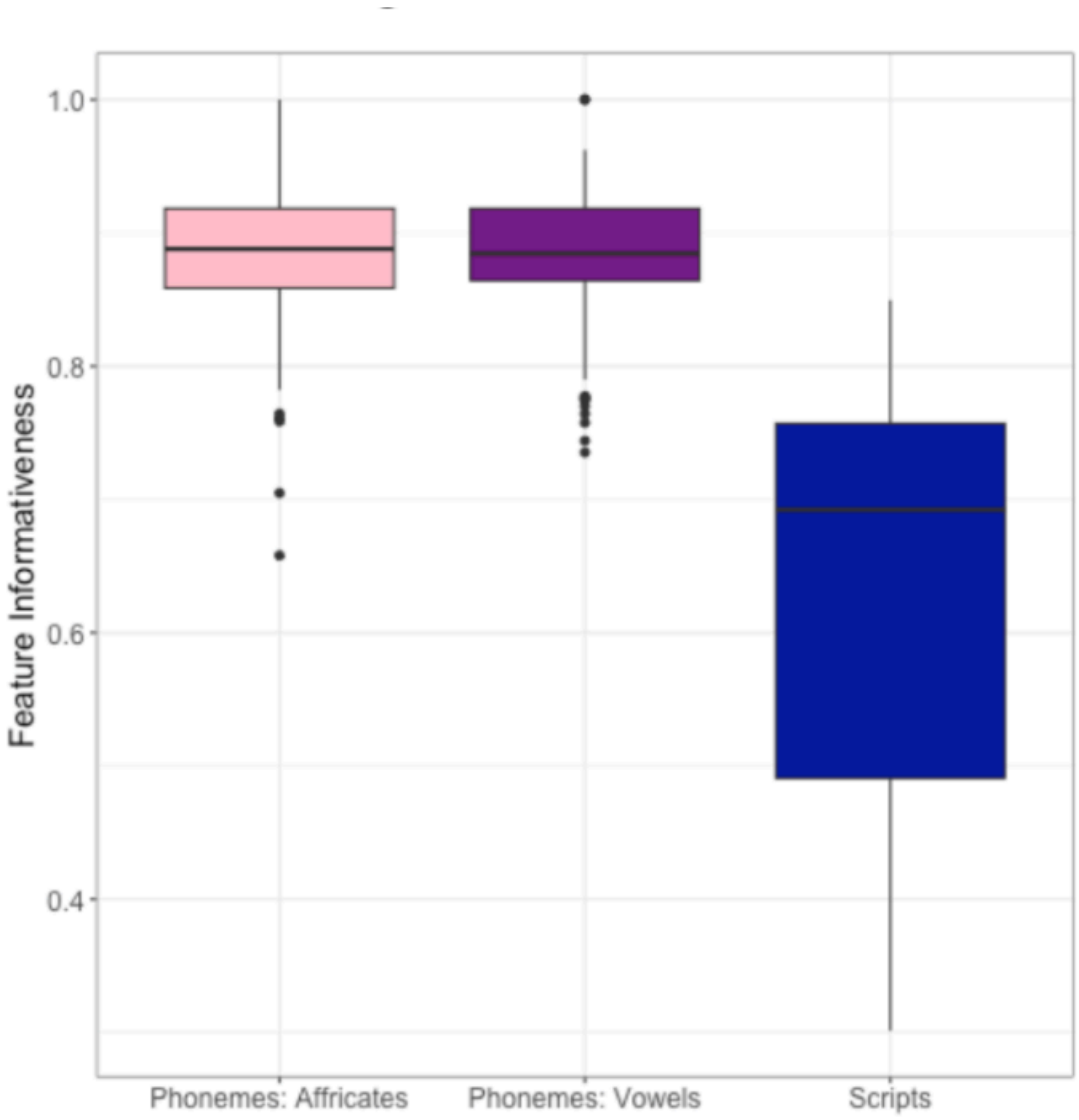
But can we try describing them with general features?



The nine general classification criteria that we applied to all scripts (left), with example application to the Buginese script. All criteria except the last two are orthogonal to one another

Prediction 2: Feature informativeness is lower in scripts

"Generalist" dataset



All predictions validated, strong effects, also when controlling for number of items (sounds/letters) in inventories / scripts.

Note: No script is completely described by the Generalist features, so we compute informativeness over all contrastive features when they apply. Likewise, for languages, we compute informativeness over all contrastive features not just the "best set".