Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The phonology of letter shapes: Feature economy and informativeness in 43 writing systems

Yoolim Kim [a,b], Marc Allassonnière-Tang [c], Helena Miton [d], Olivier Morin [e,b,*]

[a] *Linguistics Department, Carleton College, Northfield, MN, United States*
[b] *Max Planck Institute for Geoanthropology, Jena, Germany*
[c] *EA (Ecological Anthropology, UMR 7206) lab at the Muséum National d'Histoire Naturelle (MNHN) in Paris, France*
[d] *Stanford Graduate School of Business, United States*
[e] *Institut Jean Nicod (CNRS/EHESS/ENS–PSL), PSL University, Paris, France*

ARTICLE INFO

Keywords:
Writing
Reading
Letter recognition
Phonology
Combinatoriality
Feature economy

ABSTRACT

Differentiating letter shapes accurately is a core competence for any reader. Are letter shapes as distinctive as they could be? The visual shapes of letters, contrary to the phonemes of spoken languages, lack a unified description — an equivalent of the phonological features that describe most phonemes in the world's languages. Using a gamified crowdsourcing approach, we elicited thousands of letter descriptions from lay people for the sets of letter shapes (the scripts) used in 43 diverse writing systems. Using 19,591 letter classifications, contributed by 1,683 participants, who were asked to sort the letters of each script repeatedly into two groups, we extracted a sufficient number of binary classifications (features) to provide a unique description for all letters in the 43 scripts. We show that scripts, compared to phoneme inventories, use more features to produce similar sets of distinct elements. Compared to the phoneme inventories of a large sample of the world's languages dataset (the P-base dataset, collected by another team), our 43 scripts have lower feature economy (fewer symbols for a given number of features) and lower feature informativeness (a less balanced distribution of feature values). Compared to phonemes, letter shapes require more binary features for a complete description. These features are also less informative in letters than in phonemes: the chances that two random letters in a script differ on any given feature are low. Letter shapes, which have more degrees of freedom than speech sounds, use those degrees of freedom less efficiently.

Introduction

The various codes that we use to communicate and store information, from speech to writing, rely on symbols such as phonemes, words, or letters. These symbols need to be distinguished from one another for us to be able to use them. These symbols are transmitted through noisy environments where they may become distorted: they need to be sufficiently distinctive from one another in order to carry information efficiently (Dautriche et al., 2017; King, 2018; Köhler, 1987; Levy, 2008).

Distinctiveness can be achieved in two compatible ways. The first is to have symbols differ along multiple dimensions of variation. For instance, in every one of the world's languages, we find consonants that differ from one another on more than one dimension. Some consonants, like /b/ or /m/, are voiced, because producing them requires us to vibrate our vocal cords; others, like /p/, are not voiced (voiceless). Some consonants are plosives, like /p/ or /b/, being produced by letting a lot of air out in one burst. To be voiced (or voiceless) or to be a plosive (or not): such properties are phonemic features (Ladefoged, 2000). One of phonology's major discoveries is the fact that most of the phonemes of the world's languages can be described as a series of feature values: for instance the English phoneme /p/ is a voiceless bilabial plosive, /b/ a voiced bilabial plosive, etc. A relatively small set of binary features suffices to describe most of the phonemes in most of the world's languages (Mielke, 2008, Dunbar & Dupoux 2016). In spite of this, most languages possess many distinctive phonemes. This is possible because two phonemes can be identical for several features but not all.

The second path to distinctiveness is to use the features that characterize symbols informatively, meaning that the symbols exhibit as much variation as is possible on the features that describe them. Optimally, there should be an equal number of symbols exemplifying each

[*] Corresponding author.
E-mail address: olivier.morin@ens.psl.eu (O. Morin).

https://doi.org/10.1016/j.jml.2025.104620
Received 8 July 2024; Received in revised form 17 January 2025; Accepted 21 January 2025
Available online 1 February 2025

possible value that a feature can take. For instance, optimal informativeness would imply that an equal number of vowels in a language take the value "voiced" and the value "voiceless". A lack of symmetry in either direction (too many voiced relative to voiceless sounds) implies that the feature "voiced vs. voiceless" is not used to its full potential. It is not, in other words, as informative as it could be. This property has been studied and measured under various names, using various constructs, such as symmetry (Dunbar & Dupoux, 2016), combinatoriality (Changizi & Shimojo, 2005; Kirby et al., 2015; Verhoef et al., 2014; Zuidema & de Boer, 2009), or optimal dispersal (Liljencrants & Lindblom, 1972; Vaux & Samuels, 2015).

These two paths to distinctiveness — getting symbols to differ on a multitude of features, and making these features optimally informative — are compatible but different. One strategy for achieving distinctiveness consists in going far on the first path but not on the second: that is, in having a large number of features, each of them poorly informative. A set of symbols can be highly distinctive if the symbols within it vary on a very large number of features, even if each feature is not highly informative, that is to say, even if the vast majority of symbols are identical on any given feature. The opposite strategy does not go far on the first path but goes all the way on the second one. It consists in having a few features, but using them in an optimally informative way.

Spoken language is highly combinatorial

In the phonemes of spoken languages, distinctiveness is mostly achieved by means of the second strategy. Phonemes tend to possess no more than the number of features that is strictly necessary to differentiate each phoneme from all the others — a property known as feature economy or efficiency (Clements, 2003; Dunbar & Dupoux, 2016; Mackie & Mielke, 2011). Feature efficiency is not optimal in phonemes, but it is substantially higher than a range of credible random baselines would predict. Feature economy has been explained as a result of a general pressure for efficiency, including greater learnability and ease of pronunciation (Martinet, 1971). It can also be linked to a pressure for compressibility, thought to influence language evolution (Kirby et al., 2015; Verhoef et al., 2016). High feature economy minimizes the minimal description length of phonemes by making it possible to encode each phoneme as a short set of feature values.

If phoneme inventories indeed tend to possess only a small number of features, this means they can only achieve distinctiveness by making each feature as informative as possible. Do they? There is evidence that phoneme inventories are close to optimal in this respect. Optimal vowel dispersion theory, based on the premise that vowels are as widely dispersed as possible in the space defined by two formant frequencies, predicts what sounds enter vowel inventories across the world's languages relatively well, in spite of known shortcomings (Liljencrants & Lindblom, 1972). Optimal vowel dispersion can be seen as one manifestation of the more general property of symmetry, whereby phoneme inventories tend to balance the number of phonemes taking a given feature value (Dunbar & Dupoux, 2016). In keeping with previous work on dispersion, symmetry in phoneme inventories is high (though falling short of perfect optimization).

The use of a small number of features to produce a much larger number of highly distinctive symbols is a form of combinatoriality. Combinatoriality is often seen as involving combinations of discrete elements, temporally or spatially distinct: for instance, the combination of phonemes into words (Hockett, 1966; Martinet, 1971; Zuidema & de Boer, 2018). However, combinatoriality does not have to depend on the addition of discrete elements. It can rely on combinations of features, also known as signal dimensions, which are ways in which signals (or symbols) can be differentiated from one another: for instance, the combinatorial sound signals studied by Little et al. (2017) combine two dimensions, volume and duration. Visual symbols are particularly likely to use combinations of features in this way. Sign languages have a combinatorial phonology based on features such as handshape or hand position (van der Hulst & van der Kooij, 2020); heraldic emblems combine visual dimensions such as motifs and colors (Morin & Miton, 2018).

This study measures the combinatoriality of linguistic symbols, considering combinatoriality as a continuous quantity. Combinatoriality is often thought of as an all-or-nothing property. Seeing it in this way is especially useful in discussions of language evolution, since human language is exceptional in its capacity to combine meaningless symbols (phonemes) into meaningful ones (words). Much attention has thus been paid (rightly) to the question whether a given system of communication is combinatorial or not (Tamariz & Kirby, 2015; Zuidema & de Boer, 2018). Yet, a more graded view of combinatoriality can be of use too. There is more than one sense in which a system of symbols can be combinatorial (Engesser & Townsend, 2019), and even if we keep to one definition of combinatoriality, there are quantifiable differences between different systems of symbols (Galantucci et al., 2010). This paper will propose a definition of combinatoriality as consisting of two continuous dimensions. To simplify, a system of symbols is combinatorial to the extent that it uses a small number of features to generate a large number of highly distinctive symbols.

What letter shapes combinatoriality can teach us

Here, we compare the combinatoriality of phonemes with the combinatoriality of letter shapes in the scripts used by writing systems. A script is a repertoire of visual shapes used by a writing system: Latin letters are a script, as are Arabic letters, or the letter shapes that make up the Tagbanwa alphabet. A given script can be used by several different writing systems: for instance, Latin letters are a script, used by several writing systems, from Vietnamese to English.

Measuring and explaining the combinatoriality of scripts is an interesting question for at least two reasons. First, answering it helps us understand how letter shapes can be distinguished from one another. Visually similar letters are more confusable than visually dissimilar ones (Lally & Rastle, 2023; Marcet & Perea, 2017; Perea et al., 2018, 2024, Wiley et al., 2016). Having access to a repertoire of clearly distinctive letter shapes is key for cognitive and social development, given the increasing importance of literacy in our societies and the impact of letter confusion on a broad range of issues ranging from harmful reading disorders (Dehaene, 2010; Lachmann & Geyer, 2003) to lethal errors caused by doctors' bad handwriting (Bruner & Kasdan, 2001).

Second, assessing the combinatoriality of letter shapes provides a new angle from which to consider the evolution of combinatorial structure, a major puzzle of language evolution (Hockett, 1960; Little et al., 2017; Nowak et al., 1999; Zuidema & de Boer, 2009, 2018). One important mechanism for the emergence of combinatorial signals is the pressure for signals to be numerous and distinctive (Kirby & Tamariz, 2021; Scott-Phillips & Blythe, 2013). But, as we saw, combinatoriality is but one path leading to the creation of a large number of distinctive symbols. So why are spoken languages combinatorial? Three explanations are generally given.

Three possible reasons why speech is combinatorial

The first explanation is that spoken language requires the production of a large number of distinct words by a relatively simple and restricted tool: the human vocal apparatus has relatively few degrees of freedom, compared for instance to hand movements (Hockett, 1960; Little et al., 2017; Sandler et al., 2011). In this hypothesis, the pressure to produce combinatorial signals is related to the size of the space of possible signals (often called the signal space). When this space is small, signals are under greater pressure to occupy it efficiently, and thus re-use features productively. The human vocal apparatus is complex and comprises many articulator muscles, but these do not function independently of one another: instead, speech is controlled via a much smaller number of muscle groups (Pouplier, 2020; Sanguineti et al., 1997). In contrast, sign languages can avail themselves of many degrees of freedom (combining, in the case of sign language, all the possibilities of hand and arm shape, location, movement, orientation, etc. — van der Hulst & van der Kooij, 2020; Sandler, 2008). Accordingly, authors who venture to give estimates for the number of phonological features available for sign languages tend to give figures that are markedly higher than those for speech's phonological features (Mielke 2008). It should be noted, however, that we lack clear experimental evidence for a link between combinatoriality and the size of the space of possible signals. Little et al. (2017) found that increasing the dimensionality of the signal space in an artificial language evolution experiment increased combinatoriality in some respects, although the chief aim of their experiment was to explore the role of iconicity on combinatoriality (see also Verhoef et al., 2016).

Another favoured account of combinatoriality sees it as a result of loss of iconicity. Iconicity is widely thought to interfere with the evolution of combinatoriality, since combinatorial structure favors the creation of multiple random forms, while iconicity requires symbols to be motivated, thus predictable (Little et al., 2017; Roberts et al., 2015; Verhoef et al., 2016). In such accounts, loss of iconicity removes an obstacle to the evolution of combinatoriality, but it does not, on its own, explain why iconicity should evolve.

Lastly, combinatoriality has been explained as a response to the ephemeral nature of spoken sounds — speech's rapidity of fading (Galantucci et al., 2010). The fact that sounds vanish quickly after they are emitted arguably limits the possibility of forming complex holistic signals with interdependent parts; instead, it may encourage the production of symbols made of discrete, independent elements.

Each of the three factors just listed — a small signal space, low iconicity, rapidity of fading— can be plausibly related to the fact that speech is based on transient sounds. The possibility that the combinatoriality of linguistic symbols may be constrained by modality has long been entertained in the literature. Newly emerging sign languages like the Al Sayyid Sign language have been claimed to lack duality of patterning (Sandler et al., 2011), leading Little et al. (2017) to suggest that the gestural modality, with its many dimensions and large signal space, puts less pressure on hand signs to become compressible, compared to the sounds of language. The study of mature sign languages does not contradict this view, even though there is no consensus on the nature of sign languages' phonological features (or parameters) as distinct from phonemes, or on their number (van der Hulst & van der Kooij, 2020).

Letter shapes can inform this debate, since, of the three main factors thought to influence combinatoriality (a small signal space, low iconicity, rapidity of fading), two of them at least affect letter shapes quite differently than speech sounds. Regarding rapidity of fading, the difference between writing and speech is massive and straightforward: writing outlasts speech. As for signal space, the many degrees of freedom available to letters can be used to create highly distinctive shapes, without requiring a combinatorial structure. Even a highly simplified model of 2D shape generation, the LOGO model (Sablé-Meyer et al., 2021), yields an extremely high number of possible shapes—and for most of them real equivalents can be found in human cultures, showing that human hands can move freely enough to produce them. A pen or a brush's trajectory across a 2-D space can begin anywhere in the space, and change course at any point, each time in a wide range of possible directions, each bifurcation adding a new dimension to a vast space of possible signals.

Previous literature thus suggests that, with their higher signal space and permanent nature, letter shapes may not need to be as combinatorial as speech sounds; but we do not know of any attempt to answer this question. The answer is far from obvious. Some combinatorial structure is apparent in most scripts: basic shapes get reused in distinct letters (like the arch in n and h or the dot in i and j) (Changizi & Shimojo, 2005; Ladd, 2014; Meletis, 2020). And there exist a few writing systems that were explicitly (like Congolese Mandombe—Sarró, 2023) or implicitly (like Evans' syllabary for Cree) designed on the basis of combinatorial principles. Only a systematic, quantitative study can tell us to what degree the formation of letter shapes relies on combinatorial principles.

Identifying graphic features for letter shapes

We predicted that combinatoriality would be lower in scripts compared to phonemes. We used feature economy and feature informativeness to measure combinatoriality; each measure indexes a different aspect of combinatoriality. Feature economy measures the extent to which a system of symbols can generate many distinctive symbols from a small set of features; feature informativeness measures to what degree the symbols use the possibilities afforded by each feature, in order to make themselves distinctive from other symbols.

Testing our prediction requires extensive data on the features that characterize letter shapes in a broad range of writing systems. Yet research on the distinctiveness of letter shapes is overwhelmingly restricted to the study of a few important and famous scripts, in sharp contrast with phonology, where most language families can be compared systematically using standardized quantitative data. Cross-linguistic studies of letter shapes either chose to focus on a restricted set of geometrical properties (e.g. topology in Changizi & Shimojo 2005, orientation in Morin, 2018, complexity in Miton & Morin, 2021), used non-transparent manual methods that hinder replication (e.g. Changizi & Shimojo 2005), or relied on data that cannot be obtained from historical sources (e.g. stroke order in Lake et al., 2015). Attempts to build a complete descriptive vocabulary for letter shapes are restricted to one or a few writing systems and lack generalisability (e.g. Chandra et al., 2015; Primus, 2004).

This study combined the strength of all the approaches already cited by crowdsourcing a typology of letter shapes, using a white-box approach that maximises the tractability, transparency, and replicability of the visual features that characterise letter shapes. We designed Glyph, a web interface presented as a gaming applet. Thanks to extensive media coverage in several countries, it attracted a relatively diverse set of participants, overall speaking 17 different native languages (counting languages with n > 2 participants). They proposed motivated and reproducible features for letter shapes in 43 scripts. Out of those, we selected, using a decision tree algorithm, 43 sets of features sufficient to describe each script. This approach combines the granularity, tractability, and psychological plausibility of human coding with the generality and reproducibility of algorithmic approaches.

The complete data and code needed to replicate our results can be found at https://osf.io/ewp68/. This repository contains the raw data, the processed data, the results, the figures, and the code required to generate them.

Methods

This is a study of phonemic and graphemic features. For our purposes, a phonemic feature is a difference between some phonemes in a phoneme inventory and other phonemes; a graphemic feature is the same thing for the letters of a script. For example, in many languages, some consonants are sibilants while others are not; in many writing systems, some letters are symmetrical, others not. To simplify the analysis, and to align with the most complete phonological dataset at our disposal, we made the decision to consider binary features exclusively. A binary feature is a difference that separates two (and only two) sub-groups of phonemes (or letters). Although, in formal terms, the classification is created by labelling some phonemes (or letters) with the value 1 and others with the value 0, for our purposes these 0/1 labels are meaningless in themselves, and the feature is not changed by swapping all the 1s to 0s and all the 0s to 1s. The only thing that matters is the existence of two (arbitrarily labelled) groups of phonemes or letters. Such features are the basis for the two measures of combinatoriality used in our study: feature economy and feature informativeness. We first describe how these two measures were computed for phonemes, then explain how we used them to analyse letter shapes. Unless otherwise specified, phonemes and letters were analysed in the exact same way. Fig. 1 provides a graphic illustration of feature economy and informativeness, also showing the resulting measures for scripts.

P-base: An inventory of phoneme inventories

In order to compare feature economy and feature informativeness in scripts and phoneme inventories, we used a dataset of phoneme inventories that was close in format to our Glyph data, and had already been used for similar measures. The P-base dataset (Mielke, 2008) provided us with phoneme inventories for 516 languages, describing each phoneme as a set of binary feature values. P-base is strictly a dataset of phoneme inventories: it contains no information about writing systems or numerical notations. In the version we accessed (from Dunbar & Dupoux, 2016), it consists of four datasets for (1) all the phonemes in each language, (2) consonants only, (3), vowels only, (4) stops/affricates only (in this paper, stop/affricates are called “non-continuants”). Which of these phoneme inventories we considered depended on the measure we took. Of the 536 languages in P-base, we excluded four languages because they were improperly encoded in the dataset (with incorrectly formatted names). We also excluded the Taa language (also known as Southern Khoi San, ISO code nmn), due to a very large inventory size that made it a complete outlier and caused our decision tree algorithm to crash. We finally removed 15 languages because the Glottolog inventory (which we use to establish phylogenetic relationships) either did not contain them (n = 8) or did not assign them to a family (n = 7). The resulting dataset comprised 516 languages, assigned to 73 families.

Feature economy

In its simplest expression (Clements, 2003), feature economy is the ratio of the number of phonemes (or sounds, noted S) in a phoneme inventory, to the number of features (F) used to describe those phonemes: S/F. Instead of Clements’s original formula, we used Mackie & Mielke’s Relative Efficiency measure (or RE: Mackie & Mielke, 2011). Like Clement’s feature economy, Relative Efficiency compares the number of symbols in a symbol set with the number of features used to describe them, but unlike Clements’ measure, it does so in a way that controls for important artefacts linked to variation in the absolute size of S. Mackie & Mielke’s RE is given by equation (1):

$$\mathrm{RE} = 1 - \sqrt{\frac{F - Fmin}{Fmax - Fmin}} \quad (1)$$

where F is the number of features used to describe a phoneme inventory, F_{min} and F_{max} being defined as follows:

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

F_{min} is the theoretical minimum for the number of features that would be needed to describe the number of phonemes S in the inventory, assuming an optimally efficient inventory. The formula is rounded up to the highest integer because the number of features cannot be a fraction.[1]

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

F_{max} is the theoretical maximum for the number of features needed to describe the number of phonemes S in the inventory. This corresponds to using one feature for the first two symbols, then adding one feature for every additional symbol (assigning it the value 1 for the novel symbol, 0 for all other symbols).

We considered all the symbols in Glyph scripts or in the P-base inventories: i.e., all the letters shown to participants for Glyph scripts, and all the phonemes contained in the complete phoneme inventory of each P-base language. The number of symbols (letters or phonemes) was our inventory size, S. To obtain F, the total number of features needed to describe the full inventory of symbols, we used the Best Set of features for Glyph scripts (i.e., the smallest set of features capable of giving a complete description of every phoneme in that inventory—see below). To keep the P-base and Glyph data comparable, we did the same with the P-base phoneme inventories: we ran our decision tree algorithm to determine the Best Set of features for that inventory. In both cases, the size of the Best Set was our F. For 96 languages, the P-Base features were insufficient to completely describe all the phonemes in the inventory: those were discarded (remaining n = 420 languages).

Feature informativeness

Feature informativeness, a concept close to Dunbar & Dupoux’s global symmetry (Dunbar & Dupoux, 2016), measures the extent to which each feature is used in such a way as to maximize the distinctiveness of phonemes (or letters) for this particular feature. For example, if a language has six vowels, three of which are closed and three of which are open, the open/close feature is optimally informative for this language’s vowels. Feature informativeness was calculated using the Shannon entropy of the corresponding binary string (Shannon, 1948):

$$H(p) = -p \cdot log_2(p) - (1-p) \cdot log_2(1-p) \quad (4)$$

where p is the proportion of letters (or phonemes) taking the value 1 (or +, in P-base) for the feature being measured. Intuitively, informativeness is at its highest when a feature cuts an inventory of symbols in half (with 50 % symbols coded as 1 and the rest as 0). In this case, the odds that any two random symbols in the inventory present different values on this feature are maximally high. Informativeness is at its lowest when a rule singles out one symbol (with 1 symbol coded as 1 and the rest coded as 0, or vice-versa). In this case, the odds that any two symbols differ on the feature are maximally low.

When measuring feature informativeness, we want to make sure we consider only features that could logically take two values for the symbols we consider. This is not always true with P-base data: many phonological features characterize certain phoneme types but not others — for instance, being a sibilant is a feature that meaningfully differentiates affricate consonants, but not vowels or other consonants (as noted by Dunbar & Dupoux 2016). Computing the informativeness of features over all the phonemes in a language, or for overly broad categories like consonants, would result in many features not being meaningfully applicable to most sounds. For this reason, when computing feature informativeness, we only considered sets of phonemes for which we could make sure that all features took meaningful values for all phonemes. This excluded consonants and whole inventories; hence our preregistered decision to only compute informativeness on non-continuants and vowel inventories.

For each script and phoneme inventories, the Best Set of features required to describe it was computed. In a few cases (n = 30 for non-continuants, 48 for vowels), P-base features cannot give a complete description of the phoneme inventory. Those inventories were discarded.

Equation (4) was applied to each feature in the Best Set, and the measure was averaged over all features in the Best Set, yielding our measure of average feature informativeness for each inventory.

The Glyph applet

The gaming applet Glyph invited participants, who play for free and

[1] The fact that we round up $Fmin$ to the highest integer in equation (2) (following Mackie & Mielke) does not impact our results in any way. All our analyses can be replicated without rounding, by changing lines 696 and 1081 of our code to remove the ceiling() clauses, without any notable impact on the results.

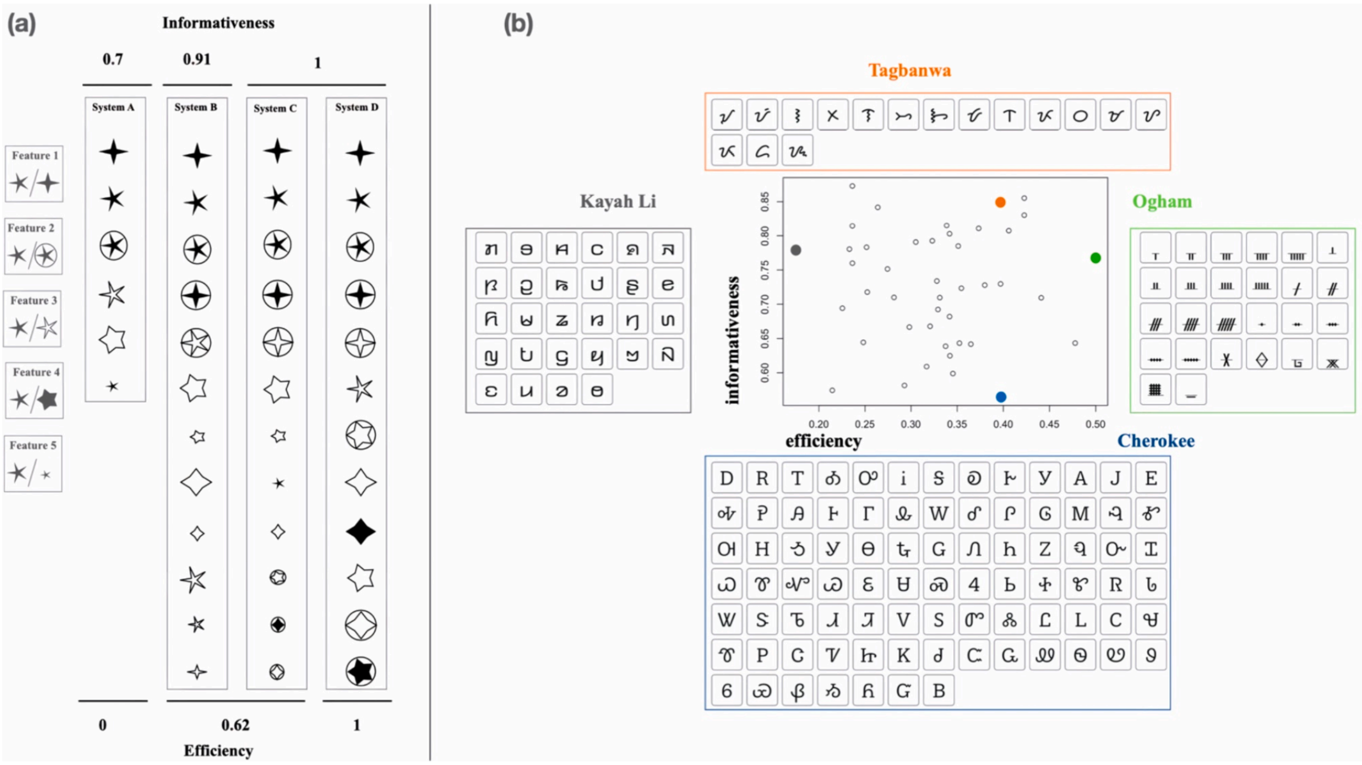
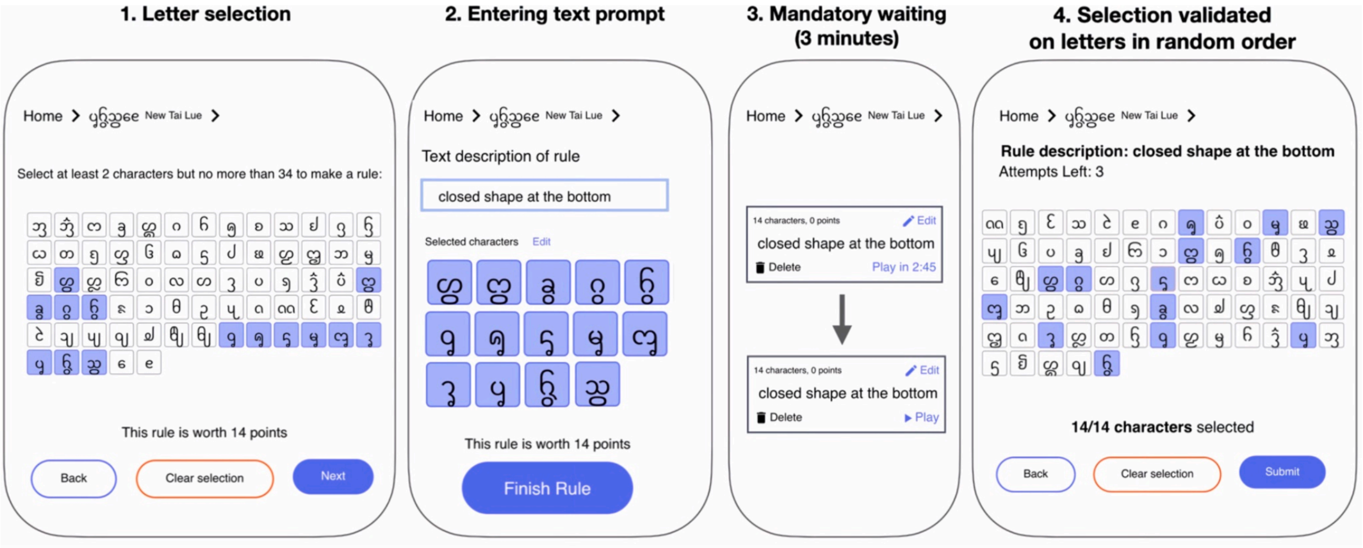


Fig. 1. Feature informativeness and feature efficiency. Panel (a): An illustration of informativeness and efficiency. The plot shows four (fictional) symbol systems A, B, C, and D, containing six or twelve symbols differing on either four or five graphic features (shown on the left). In the most informative systems (C and D), each feature value is instantiated in exactly half of the symbols, this balance being less perfect in the less informative systems. The most efficient system (D) uses only four features to generate 12 different symbols: this system is perfectly efficient since it is impossible to use fewer features and generate as many distinct symbols. The least efficient system (A) uses five features to generate six symbols, the lowest possible number of symbols that can be generated from five features. As systems B and C show, it is possible for two systems to be equally efficient but differ in feature informativeness. **Panel (b): Feature efficiency and feature informativeness in 43 scripts**, with four sample scripts illustrating extremes of both measures. Kayah Li and Ogham have roughly equally informative features, but differ starkly in their efficiency: a very small number of features suffices to describe Ogham letters, whereas Kayah Li letters differ in many more ways. This pattern is reversed for Tagbanwa and Cherokee: those have roughly equal feature efficiency, but the graphic features of Tagbanwa are much more informative than those of Cherokee. As the plot shows, feature efficiency and feature informativeness across scripts are almost perfectly uncorrelated (Spearman’s rho = 0.00).

for fun, to devise letter shape classifications for 43 scripts representing a broad range of linguistic, semiological, and historical variation — from alphabets to syllabaries and from Indo-European languages to Sino-Tibetan. These writing systems were selected based on the following criteria (see supplementary materials for more detail on script selection). First, relative obscurity. We wanted participants to classify letter shapes based purely on their visual appearance, which they could not do if they were literate in the script in question. We therefore chose to exclude most of the world’s widely known scripts (including Latin, Cyrillic, Arabic, etc., the Greek script being used only for the tutorial phase). A few relatively well-known scripts were included, such as Hangul, due to their theoretical importance (Hangul being of interest as a *de novo* invention and as a putative featural script). We systematically removed from analysis all the classifications proposed by participants literate in the script they sorted. Second, we only selected scripts that could be entirely sorted manually by a human being in a reasonable amount of time, excluding all very large scripts such as Chinese logographic symbols, the Cree syllabary, and others. Lastly, we strove to build a dataset of scripts that were as phylogenetically independent from one another as was possible, in order to maximise the independence of data points. No two scripts in our dataset have a direct ancestor–descendant relationship, and we minimized the number of scripts having a common ancestor in their genealogy. We also prioritised *de novo* inventions such as Hangul, Shavian, or Bamum. Such inventions are visually quite distinct from surrounding scripts (even though influences can never be completely excluded), unlike standard scripts, which usually bear clear marks of ancestral influences. A few *de novo* scripts are inventions that never came into frequent use (e.g. Shavian, an unrealised attempt at replacing the latin script for writing English). Although they may appear anecdotal, such scripts provide us with a unique window into the genesis of letter shapes and increase the diversity of datasets otherwise dominated by the descendants of a few big scripts (see Kelly et al., 2021 on the importance of *de novo* inventions for studying script evolution; also Roberts & Galantucci 2012 make a parallel case for *de novo* evolving languages).

After a brief tutorial explaining the basic principle of the game, Glyph participants were invited to produce two classifications for the Greek script, as a way to ensure they understood how the applet worked. After this, they were shown the full list of scripts available for playing, among which they were free to pick a choice. (The supplementary materials include a full description of the applet from the user’s point of view, with screen captures.) To earn points, participants have to select a script and propose a binary classification of letters for that script (Fig. 2). Each selected letter is marked as having value 1 for the relevant classification, the other letters having the value 0. To form a classification, a player may select as few as two letters and as many as half the total number of letters in the script.

Once done, the player is asked to gloss the classification they have proposed with a short text. Entering this text starts a three-minute countdown, at the end of which the player is allowed to try and validate the classification they proposed earlier. To validate a classification, the player is invited to replicate it on the exact same letters, presented in a different, randomized order. The text they wrote at the previous step is presented as a prompt. The player must reproduce the exact same



1. Letter selection

Home > New Tai Lue >

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

This rule is worth 14 points

Back | Clear selection | Next

2. Entering text prompt

Home > New Tai Lue >

Text description of rule

closed shape at the bottom

Selected characters Edit

This rule is worth 14 points

Finish Rule

3. Mandatory waiting (3 minutes)

Home > New Tai Lue >

14 characters, 0 points Edit
closed shape at the bottom
Delete Play in 2:45

14 characters, 0 points Edit
closed shape at the bottom
Delete ▶ Play

4. Selection validated on letters in random order

Home > New Tai Lue >

Rule description: closed shape at the bottom
Attempts Left: 3

14/14 characters selected

Back | Clear selection | Submit

Fig. 2. The production of letter classifications in Glyph. Players first propose a classification of the letters of a script into two groups, according to some criterion, which they have to write down as a text prompt. After a mandatory waiting period of three minutes, they are asked to replicate their classification exactly, on the same letters presented in a different, random order, based on the text prompt they wrote.

classification for all the letters in the script. They get three trials for this; after three failed trials, a classification can no longer be validated. A validated classification earns its creator a number of points equal to the number of letters selected by the classification. The number of points is doubled if the classification is unique (that is, it has not yet been proposed by any previous player).

To make sure that classifications are based on visual features alone, and not on a knowledge of the underlying language, we ask players to disclose all the languages they are literate in, and we automatically remove all their classifications for the relevant scripts. We also went through the Best Set classifications manually to remove all classifications not based on letter shapes alone.

As of writing this, around 5,000 players registered on the applet and produced around 100,000 classifications for our scripts, c. 30,000 of them being unique. Following a preregistered cut-off decision, we only consider in this study the data obtained between February the 4th and August 20th, 2022: 44,911 classifications, 19,591 of them unique, from 1,683 players. For each individual script, between 169 and 1,660 distinct classifications were proposed. The Glyph data yields crowd-sourced typology of letter shapes capable of identifying each individual letter in all 43 scripts (Fig. 3).

Glyph-produced features

To match the phonologists' terminology, the binary classifications produced by Glyph players are called, here, "features". Glyph players

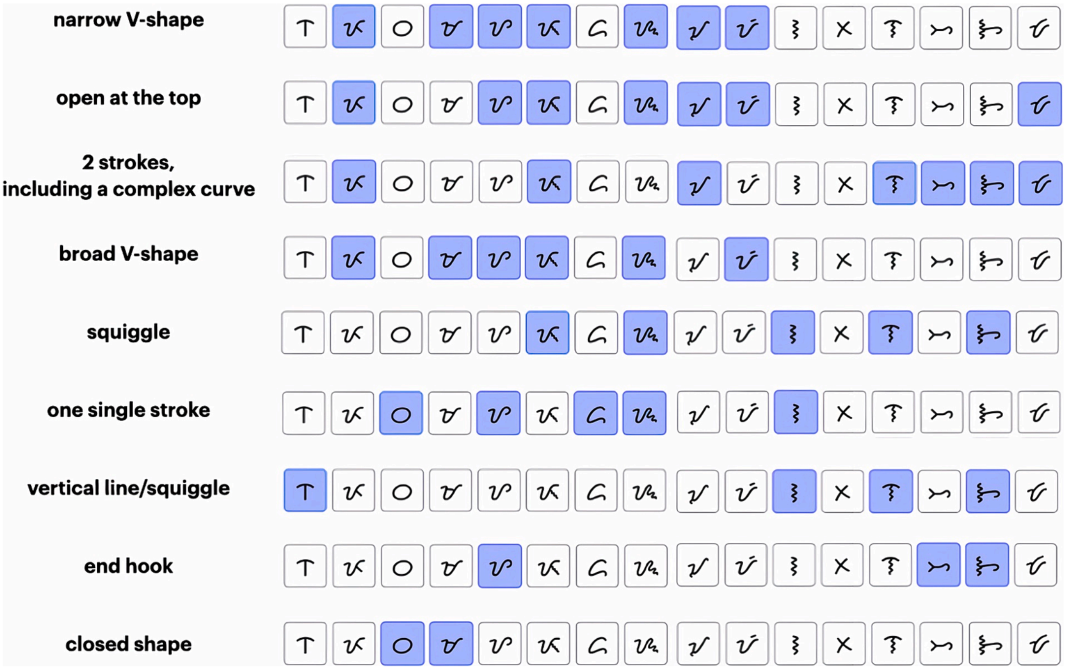


Fig. 3. The Best Set of classifications, for the letters of the Tagbanwa alphabet. Blue squares correspond to 1 values, white squares to 0 values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

labeled each of their classifications with a verbal prompt describing the letters that the participant selected by clicking on them (e.g. “All symmetrical letters”, “All letters without a middle bar”, etc.). Our analyses in this paper focus on the underlying classifications. These classifications are, from the point of view of feature informativeness and efficiency, perfectly reversible. For the set of four letters {A,B,C,D}, the classification (A = 0, B = 0, C = 1, D = 1) is considered identical to the classification (A = 1, B = 1, C = 0, D = 0). In linguistic parlance, the features are considered (for the purposes of our analyses) as having no marked value. As a consequence, the players’ choices are not limited by the fact that we forced them to select at most half the letters in a script. Suppose for instance that a player wants to create a rule based on the presence of a vertical line, for a script of 20 letters where all letters except two have a vertical line. It will not be possible for them to pick all 18 letters with a vertical line, but they can pick the two letters without —which amounts to the exact same classification. The fact that many features are described negatively shows that Glyph players clearly understood this simple principle.

In all these respects, Glyph features are similar to the format of the P-base data used in this paper’s analysis: P-base features are also analysed as simple binary classifications. Glyph features differ in one respect, though: the points system encouraged participants to produce informative classifications (i.e. classifications that come as close as possible to splitting the set of letters evenly) and we even made it impossible to single out one letter for a classification. One of our study’s goals was to see how far we could go in describing letter shapes using a few binary features. This objective requires features to be as informative as possible, and it rules out singleton classifications which are, informationally speaking, trivial — since they merely single out one letter against the rest of the script. These decisions bias our methods and our data *against* the hypotheses and results presented here (a consequence we consider in the Discussion).

The best sets of features

We need to determine on which features exactly our two measures, feature economy and feature informativeness, were to be computed. Since one of our datasets (Glyph) contains very large numbers of features for each system, some of them of little interest, we rule out computing each measure on all features. Also, since feature economy is explicitly a measure of the smallest number of features needed to fully describe a phoneme inventory (or a script), we decided to find this set, which we call the “Best Set”, and make it the basis for our measurements. The following method was used on both P-base and Glyph features. We used a decision tree algorithm (rpart package in R, closely inspired by the CART algorithm, Therneau et al., 2023) to pick, out of all the features proposed for a given script or phoneme inventory, the smallest set of features capable of giving a complete description of the script or phoneme inventory. A set of features completely describes a phoneme inventory when every letter in the script corresponds to a different value for the whole set of features—in other words applying the features to each letter (or phoneme) produces a different binary string for every letter (or phoneme). These Best Sets of features form the basis of our analyses for both letters and phonemes (Fig. 4).

Validating Glyph features as measures of similarity

We represent each of our 43 scripts as a set of visual features with value 0 or 1, each letter being identified as a unique binary string representing feature values. The distance between any two letters can then be computed as the Hamming distance between the corresponding binary strings.

We validated these letter descriptions by showing that they predict participants’ judgments of similarity between letter pairs. We chose ten Glyph scripts, picked six letter pairs for each script, and ran a preregistered validation experiment, asking 180 US participants recruited from Prolific to rate each of the 60 pairs for visual similarity, on a Likert scale. We found a substantial correlation between the average similarity rating given by participants to a given letter pair, and the Hamming distance between the two letters according to Glyph features (Spearman’s rho = 0.75).

Controlling for the non-independence of data points

We complemented P-base data with information on language families from Glottolog (Hammarström et al., 2023) to avoid giving too much statistical weight to over-represented families, and to account for common ancestry (Mace & Holden, 2005). The 43 scripts included in the Glyph dataset were selected to form a representative sample of writing systems reflecting their linguistic, geographic, and typological diversity. They were selected to avoid any pair of directly related scripts, and to include a large number of *de novo* inventions. This mitigates the need to control for ancestry (and makes it difficult to do this, should we want to). Therefore, we assigned a dummy family to each script, basically considering each script to form its own family.

The methods and measurements used and the sets of languages and scripts studied, were preregistered in advance of analysis (see supp. mat.). However, the results shown here are part of a broader set of preregistered studies, not all of which are shown here, and the theoretical interpretation of the results is partly post-hoc. The supplementary materials give a complete summary of the preregistration documents.

This research has received the approval of the Ethik-Kommission affiliated with Universitatsklinikum (Ethics Committee) at Friedrich Schiller University Jena (approval number 2021–2118-Bef). Data collection and dissemination was carried out in line with the European Union’s GDPR. Participants consented to the study’s terms and conditions.

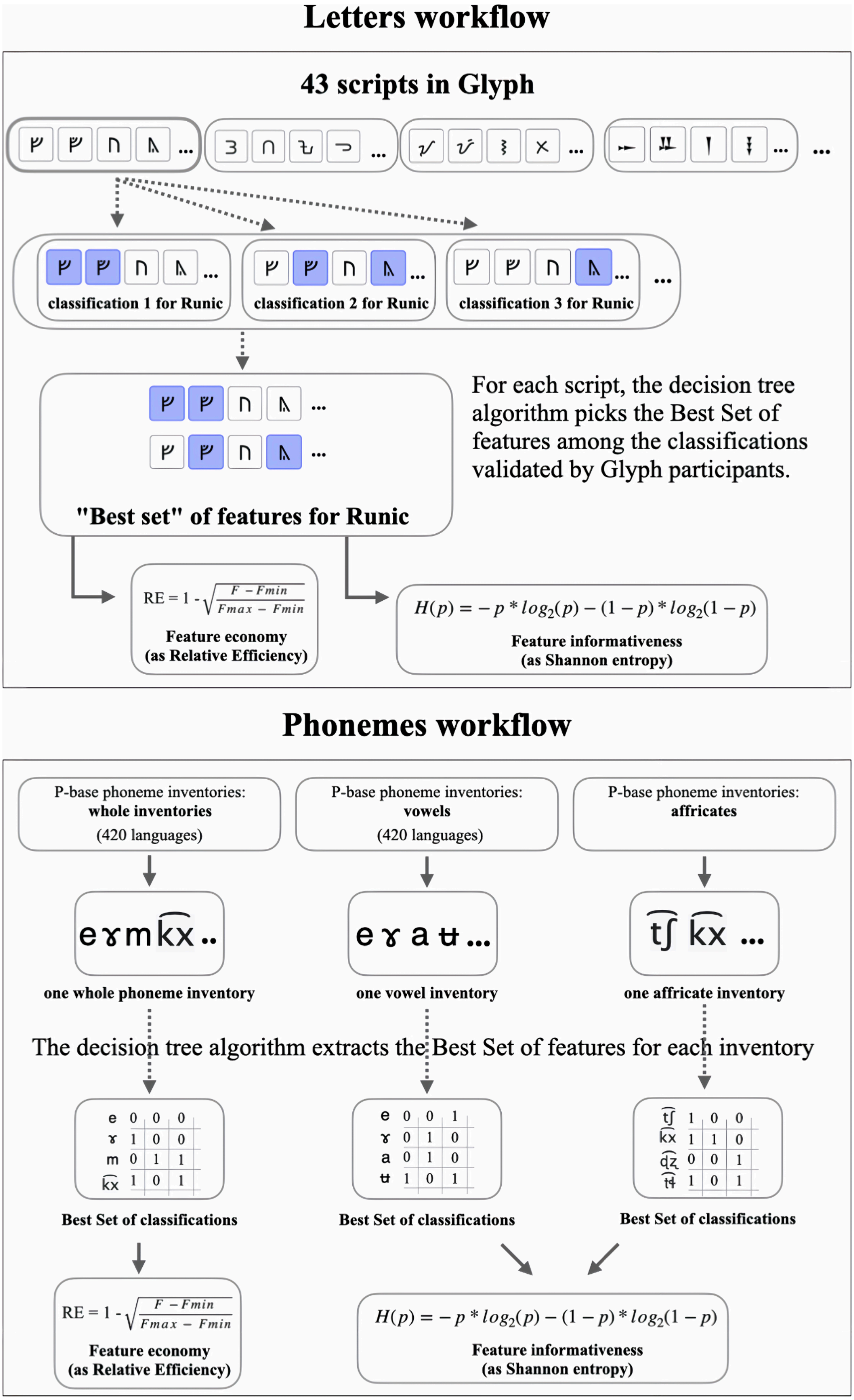
Results

Scripts are weakly combinatorial, compared to phoneme inventories

The number of features required to provide a complete description of each script rises linearly with the number of letters in the script (Pearson’s r = 0.91). Regressing the number of letters in a script over the number of features required to describe it, we find a coefficient of 2.07 (95 % CI: 1.8/2.3). In other words, for every addition of two letters in a script, one new feature is needed to describe the script. In theory, if scripts were organized in a maximally efficient way, seven orthogonal features would be more than sufficient to describe any script of size 85 or less (85 is our maximum script size). In practice, all scripts require about half as many features as they have letters. Phonemic inventories are much more combinatorial, the regression coefficient (predicting the number of phonemes in an inventory over the number of features required to describe it) being 3.7 (95 % CI: 3.4/3.9). That is almost double the coefficient for scripts. Put differently, the number of extra phonemes that would have to be added to a phoneme inventory to justify the addition of one extra feature is almost twice higher for phoneme inventories compared to scripts.

Feature economy is lower in scripts

We compared two mixed effects models, each designed to predict the feature economy of a phoneme inventory or script. Each time, we compared two models, a null model predicting each data point’s feature economy based on language family (or script) alone, and a test model similar to the null but for the addition of our variable of interest: the type variable, a categorical variable stating whether the data point is a script or a phoneme inventory. The test model was more informative than the null model (Δ_{AIC} = 20), and the weight attached by the model to the “type = script” variable is negative (β = -0.08, 95 % CI: −0.10/-0.05). Feature economy is lower in scripts compared to phoneme inventories


Letters workflow
43 scripts in Glyph
classification 1 for Runic
classification 2 for Runic
classification 3 for Runic
For each script, the decision tree algorithm picks the Best Set of features among the classifications validated by Glyph participants.
"Best set" of features for Runic
$RE = 1 - \sqrt{\frac{F - Fmin}{Fmax - Fmin}}$
Feature economy
(as Relative Efficiency)
$H(p) = -p * log_2(p) - (1-p) * log_2(1-p)$
Feature informativeness
(as Shannon entropy)
Phonemes workflow
P-base phoneme inventories:
whole inventories
(420 languages)
P-base phoneme inventories:
vowels
(420 languages)
P-base phoneme inventories:
affricates
e ɤ m k͡x ..
e ɤ a ʉ ...
t͡ʃ k͡x ...
one whole phoneme inventory
one vowel inventory
one affricate inventory
The decision tree algorithm extracts the Best Set of features for each inventory
Best Set of classifications
Best Set of classifications
Best Set of classifications
$RE = 1 - \sqrt{\frac{F - Fmin}{Fmax - Fmin}}$
Feature economy
(as Relative Efficiency)
$H(p) = -p * log_2(p) - (1-p) * log_2(1-p)$
Feature informativeness
(as Shannon entropy)

e	0	0	0
ɤ	1	0	0
m	0	1	1
k͡x	1	0	1

e	0	0	1
ɤ	0	1	0
a	0	1	0
ʉ	1	0	1

t͡ʃ	1	0	0
k͡x	1	1	0
d͡ʑ	0	0	1
t͡ɬ	1	0	1

Fig. 4. A schematic presentation of the study's workflow for analyzing phoneme inventories (top) and scripts (bottom).

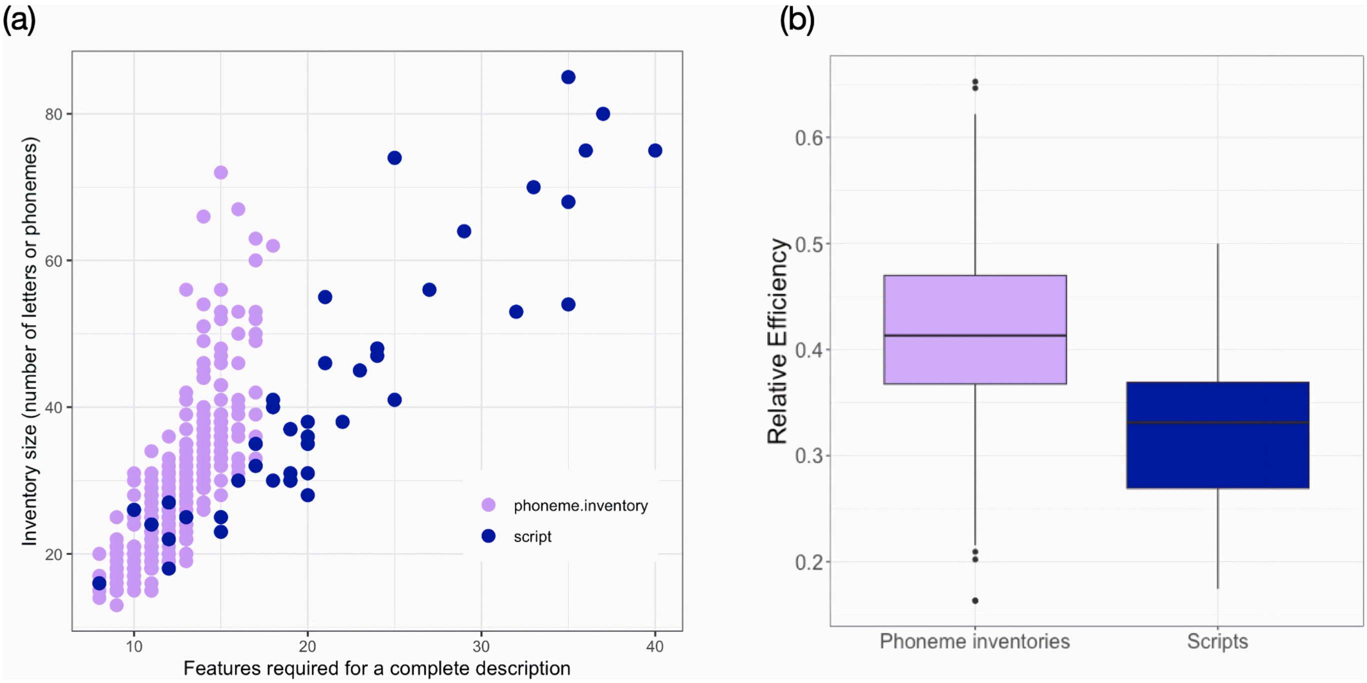


Fig. 5. Combinatoriality in scripts and phoneme inventories. (a) The inventory size (number of phonemes or letters) for phoneme inventories or letters (y axis), plotted against the number of features required for a complete description of the letters or phonemes (x axis). Scripts require more features per item compared to letters. **(b)** Feature economy, measured as Mackie & Mielke's relative efficiency, in phoneme inventories (n = 420) and scripts (n = 43). The boxes in the boxplot represent the two middle quartiles.

(Fig. 5). Adding a control for inventory size, i.e. the number of letters (or phonemes) in a script (or phoneme inventory), produces more informative models but does not change this effect (see the open code, section 6.2; https://osf.io/ewp68/ > 2. Code).

Letter features are less informative than phonemic features

We measured the average informativeness of features in scripts compared to two types of phoneme inventories — vowels and non-

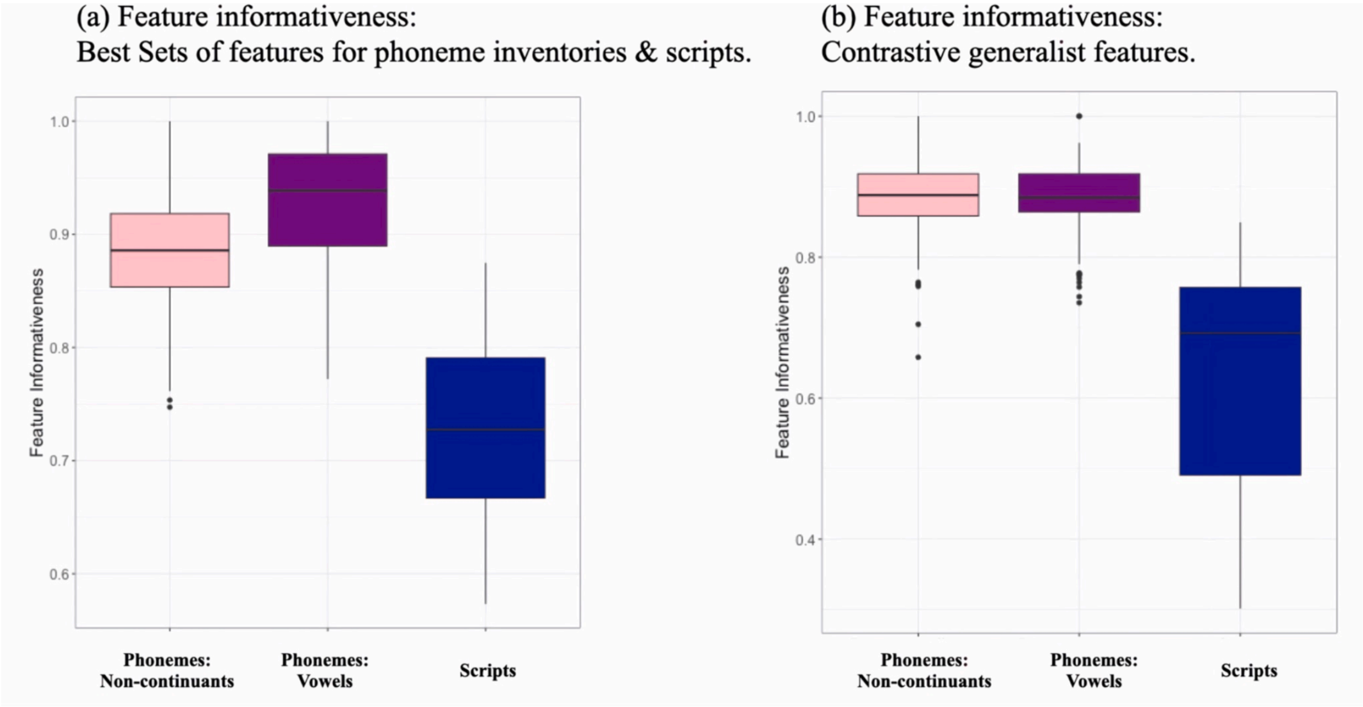


Fig. 6. Feature informativeness in non-continuants or vowels inventories (n = 486 / 468) and in scripts (n = 43). (a) Here, feature informativeness is computed over the Best Set of features (the smallest set of features capable of describing all the symbols in a set of symbols) both for phoneme inventories and for scripts. **(b)** The same comparison, using only generalist features, which are in principle applicable to all scripts (our data on phoneme inventories always use such features). Here, feature informativeness is computed over all the contrastive features (i.e. features that take more than one value) for a given phoneme inventory or script. Boxes represent the two middle quartiles.

continuants. The data points were individual scripts or phoneme inventories. Each time, we compared two models, a null model predicting each data point's feature informativeness based on family alone, and a test model similar to the null but for the addition of the type variable. As predicted, the test model was consistently more informative than the null model (Δ_{AIC} = 104 for the comparison with non-continuants, 147 for the comparison with vowels) and the weight associated with the "type = script" variable was negative (comparison with non-continuants: β = -0.15, 95 % CI: −0.15/-0.13; with vowels: β = -0.19, 95 % CI: −0.19/-0.17). Adding a control for inventory size, i.e. the number of letters (or phonemes) in a script (or phoneme inventory), produces more informative models but does not change this effect. The null model with family and number of items is outperformed by the test model (family + inventory size + type) for both comparisons — with non-continuants or vowels (Δ_{AIC} = 20 and 32 respectively), with a negative effect for "type = script" (β = -0.07, 95 % CI: −0.07/-0.04, and β = -0.09, 95 % CI: −0.09/-0.07). Graphic features are less informative than phonemic features (Fig. 6**a.**).

Our result on feature informativeness replicates with a general classification of letter shapes

One major difference between phonemes (as studied in P-base) and letter shapes (as studied here) is the fact that all phoneme inventories are described using the same 24 features, whereas the set of graphic features used to describe scripts is unique to each script. To address this, we sought to replicate our results on informativeness using only graphic features that could apply to a broad range of scripts. We picked nine generalist criteria chosen among the most common descriptions present in the Best Sets of classifications across scripts (Fig. 7). We asked two independent coders to apply each of these criteria to all the scripts. For each script, we retained a criterion if the two coders agreed over its application to that script (Cohen's kappa > 0.61), which was true in 94 % of cases. To keep the comparison equal between scripts and phoneme inventories, we disregarded cases where the application of a criterion to a script yielded only 0 or 1 values, with no contrasting values (e.g., all letters are asymmetrical, no letter contains crossing lines, etc.) (21 % of remaining cases). Our nine generalist criteria are never sufficient to describe all the letters in any script, therefore we cannot use it to test any hypothesis related to feature economy.

We replicated our results concerning feature informativeness using these general features. This time, script informativeness was calculated over all generalist features (not on the Best Set). Symmetrically, for the P-base phoneme inventories, feature informativeness was computed for all the features that were given two contrastive values for the relevant phoneme inventory (rather than computing informativeness on the Best Set of features for this inventory). We compared a null model predicting each data point's feature informativeness based on family alone, and a test model similar to the null but for the addition of the type variable. As predicted, the test model was consistently more informative than the null model (Δ_{AIC} = 91 for the comparison with non-continuants, 90 for the comparison with vowels) and the weight associated with the type variable was negative (comparison with non-continuants: β = -0.24, 95 % CI −0.20/-0.27; with vowels: β = -0.24, 95 % CI −0.20/-0.28). Adding a control for inventory size, i.e. the number of letters (or phonemes) in a script (or phoneme inventory) does not change this general pattern of results (see see the open code, section 5.3; https://osf.io/ewp68/ > 2. Code).

We also replicated this result using only seven general criteria that are orthogonal to one another, meaning that the value a letter takes for one of these criteria is logically independent of its value on the other six. For instance, whether a letter contains an enclosed space or not does not impact its shape with regard to the other criteria. To get these seven criteria, we removed two criteria that overlapped with one other criterion that could be applied consistently to more scripts: the criterion "symmetry" (which our participants applied less consistently than the criterion "vertical symmetry"), and the criterion "separate parts" (which is not orthogonal with the criterion "can be drawn in one stroke" and

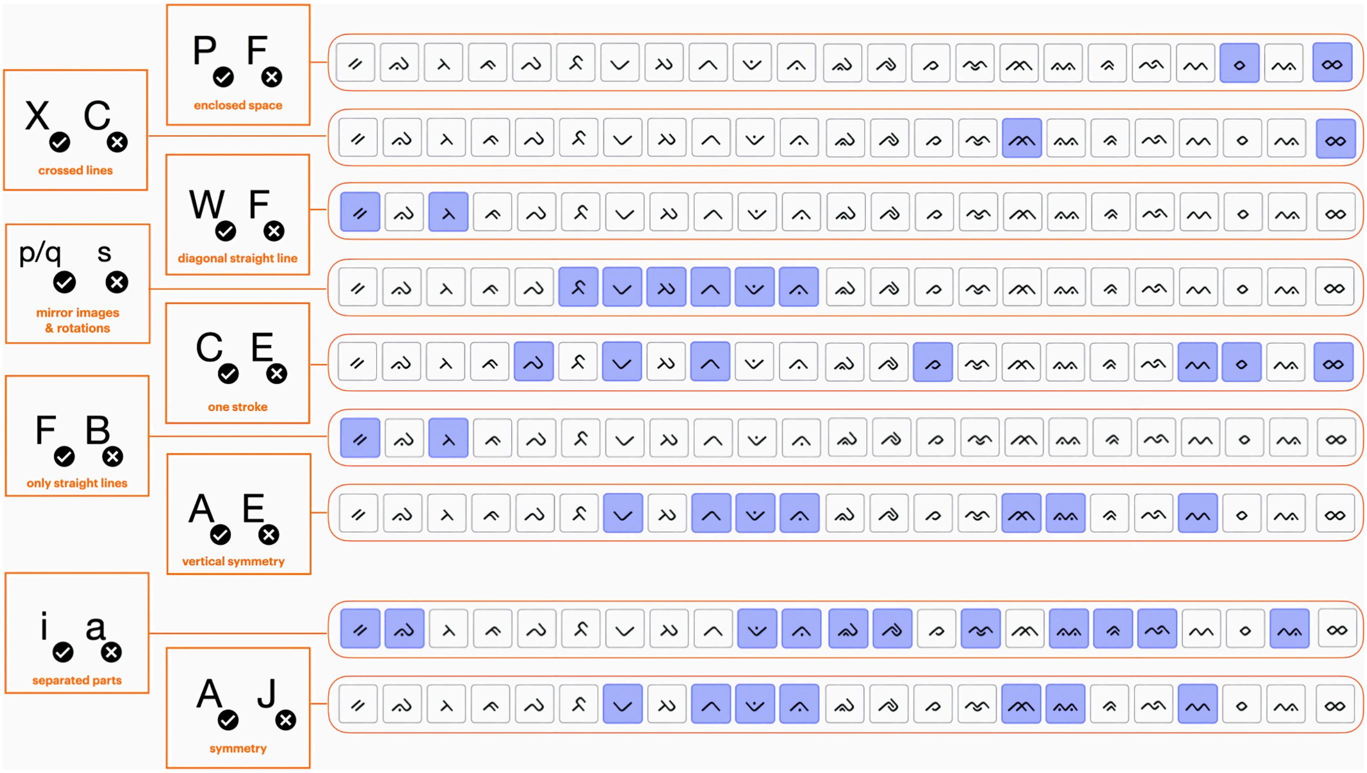


Fig. 7. The nine general classification criteria that we applied to all scripts (left), with example application to the Buginese script. All criteria except the last two are orthogonal to one another ("separated parts" is not orthogonal with "one stroke", and "symmetry" is not orthogonal to "vertical symmetry").

applies to fewer scripts). Doing this does not change our results.

Discussion

Letter shapes only have limited combinatoriality

We used a unique dataset built from crowdsourcing letter descriptions across 43 writing systems to produce a comprehensive typology of letter shapes for these diverse scripts. We managed to extract, from the classifications contributed by thousands of Glyph players, enough features to provide a unique description of all letters in all the writing systems we studied. These features were validated within participants; they are extensively glossed in natural language; and they predict outside participants' judgments of similarity between letters. The Glyph dataset improves upon previous attempts at establishing cross-linguistic typologies of letter shapes in being psychologically valid, transparent to a human reader, powerful enough to describe all letter shapes, and free of theoretical commitments.

This white-box approach to the typology of letter shapes offers results that are consistent with previous findings by Changizi & Shimojo (2005), the only other comprehensive study to date addressing the combinatoriality of letter shapes with a broad comparative dataset. Consistent with the results they obtained with extremely different methods, we find scripts to be combinatorial only to a limited degree. Scripts are even less combinatorial in our estimate than according to theirs: they showed that the number of stroke types T used by letters in a script increases with script size S according to the function $S = T^{3/2}$. In contrast, we find that the number of features required to completely describe a script increases linearly with the number of features in the script, the relationship being consistent with the addition of one feature for every two letters. We also address, for the first time, another aspect of combinatoriality, the informativeness of features, and find it to be, here again, quite low. Most importantly, we assess the combinatoriality of scripts against a plausible comparison point, the combinatoriality of phoneme inventories, finding phonemes to be more combinatorial than letters on our two metrics, feature economy and feature informativeness. These two measures being completely uncorrelated as far as scripts are concerned, the two tests are independent of each other: each provides an independent confirmation of our hypothesis.

Our dataset contains two scripts of particular interest, Hangul (for Korean) and Shavian (for English). Those share two unusual properties: they were created from scratch (as opposed to evolving from a clearly identifiable predecessor) and their creators intended them to reflect phonemic features. Letter shapes in Hangul and Shavian are not meant simply to encode phonemes, but also to reflect sub-phonemic properties such as place of articulation, voicing, etc. They succeed in doing so to the extent that visual similarity in these writing systems correlates with phonemic similarity of the encoded sounds (Jee et al., 2022). If letter shapes in these systems truly mirrored the featural organization of the phoneme inventories of their target languages (Korean or English), we would expect these scripts to have relatively high feature economy and informativeness, just like the phonemes that they represent. Yet, neither Hangul nor Shavian rank very high for either feature economy or informativeness, even though the phonemes that they encode (those of Korean and English, respectively) have regular (that is to say, high) feature economy and informativeness.[2] This could be for several compatible reasons. First, neither Hangul nor Shavian reflect phonemic features with perfect consistency (Sohn, 2001). Second, the graphic features with which these scripts encode phonemic features may not be sufficiently salient, visually speaking. Shavian makes extensive use of mirror inversions and rotations, which are notoriously difficult to perceive (Fernandes & Leite, 2017).

Limitations

An obvious limitation of this study lies in the comparison of two datasets of different provenance, analyzed using vastly different (yet comparable) methods. Discrepancies between datasets are an intrinsic part of any comparative research, but our use of a crowdsourcing method to describe letter shapes raises specific issues. The P-base dataset is the result of decades of work by phonologists and ethnolinguists equipped with a standardized toolbox containing, among other things, the international phonetic alphabet and elaborate theories of phonological features. The effort of hundreds of committed Glyph participants cannot even approximate the scale of that scientific enterprise. This could bias our results, to the extent that there are some highly informative ways to classify letter shapes that Glyph participants failed to notice, for lack of expertise. How much room is there to improve upon Glyph classifications?

Our data give us a clue (see supp. mat.). If we consider the evolution of the Glyph dataset, ordering the classifications proposed by participants chronologically from the applet's opening day to the end of data collection time, we can see how much progress was made at each time step. It took relatively little time to reach the stage where Glyph classifications can uniquely identify every letter in every script: 97 % of letters were uniquely identified already by the first 20 % classifications. The script descriptions that we can extract from the Glyph dataset do not become more efficient after that point: the feature economy of scripts (measured as Relative Efficiency) does not increase after the first 20 % classifications—if anything, it decreases. This suggests that the first players picked most of the low-hanging fruits, and that the rate of progress should continue to slow down: room for improvement appears limited, unless a scientific breakthrough occurs.

The comparison between Glyph and P-base also contains biases that go *against* our two hypotheses, and thus strengthen our results. Regarding feature informativeness, we incentivized Glyph participants to produce informative classifications, in two ways: the number of points earned for each classification was a direct function of the classification's informativeness (and participants knew this); furthermore, classifications could not be validated unless they singled out at least two letters. This imposes a minimal informativeness threshold on all Glyph features, in contrast with P-base features, which can and occasionally do characterize one phoneme only.

Regarding feature efficiency, the decision tree algorithm that computes the Best Set of features is more efficient, that is to say, more likely to yield parsimonious descriptions, if it has many different features to choose from (all else being equal). When analyzing the P-base data, we only allow our algorithm to access P-base's 24 features; when analyzing Glyph data, however, our algorithm can usually access hundreds of classifications for each script it considers, all of them tailor-made for the script at hand. In theory, this could have made it easier for our algorithm to generate parsimonious descriptions for scripts compared to phoneme inventories; in practice, it did not, because Glyph participants could not propose sufficiently distinct and informative classifications for letter shapes.

The building blocks of letter shapes

The dataset we assembled treats letters as combinations of discrete features (taking binary values), in line with research in phonology. We are not committed to the view that letters are mentally processed as sets of discrete features, but we note that this claim has been backed by substantial evidence (Grainger et al., 2008; Pelli et al., 2006). Our data is compatible with Pelli's proposal that even relatively complex letters may be encoded with a small number of features (Pelli et al., 2006).

[2] Feature informativeness for Shavian letter shapes: 0.45; for English phonemes, 0.85 (vowels) and 0.91 (affricates). Feature informativeness for Hangul letter shapes: 0.5; for Korean phonemes, 0.89 (vowels) and 0.84 (affricates). Feature efficiency for Shavian letter shapes: 0.33; for English phonemes: 0.44; for Hangul letter shapes: 0.40; for Korean phonemes: 0.38.

When we consider, for each letter in each script, the number of visual features that suffice to distinguish this particular letter from all other letters in the script, the average result ranges from 4.2 to 9.3, with an average value of 6.5 across all scripts, consistent with Pelli et al.'s claim that letter identification is mediated by the recognition of about 7 visual features.

Our massively crowdsourced dataset allows us to study letter shapes in an entirely novel way, combining the granularity of script-specific approaches with the power of comparative datasets, and improving upon past comparative studies by using classifications that are open, reproducible, and uninfluenced by theoretical commitments. The measure of letter distinctiveness that we derive from this data improves upon similarity measures that are either not validated against human judgments (e.g. Antic & Altmann, 2005; Han et al., 2022) or show a poorer fit with them (e.g. Jee et al., 2022). Our dataset provides a set of plausible candidates to start investigating the basic components of letter identification.

Conclusion

Our study of letter shape combinatoriality in a diverse sample of writing systems suggests written symbols use their features less efficiently than speech sounds do. This is consistent with the view that the combinatoriality of spoken language is, in part, a consequence of speech's rapidity of fading and the relatively small size of its signal space. At the same time, the fact that letter shapes show a substantial, albeit lower level of combinatoriality suggests that these factors cannot (jointly or separately) provide a full explanation for why human symbols combine smaller sets of features to produce bigger sets of distinctive symbols. Only experimental studies can really tease apart the various mechanisms underlying the evolution of combinatorial communication; what this observational study did was to provide a rigorous and general estimate for the efficiency and informativeness of graphic features. The logical next step is to try and identify a series of features that, together, could give a complete description of letter shapes in the world's writing systems; but, if our results provide any cue, the task will be demanding.

Funding

This research and the glyph website were funded by the Max Planck Society. We also acknowledge the support of the ANR grant ComCogMean ANR-23-CE28-0016 as well as the EUR FrontCog grant ANR-17-EURE-0017 and ANR-10-IDEX-0001-02, and the PSL SCRIPTA program.

CRediT authorship contribution statement

Yoolim Kim: Writing – review & editing, Software, Project administration, Data curation, Conceptualization. **Marc Allassonnière-Tang:** Writing – review & editing, Formal analysis. **Helena Miton:** Writing – review & editing, Validation, Methodology, Conceptualization. **Olivier Morin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to M. Deleron, A. Depierreux, A. Dietrich, A. Felk, K. Hamilton, S. Kayankit, A. Koshevoy, L. Kurnosova, S. Lee, D. Martini, T. F. Mueller, M. Myung, P. Nole, B. Pavlek, H.-G. Sell, O. Sobchuk, S. Steele-Dadzie, A. Thomas, A. Tsuchiya, R. Urbanek, E. Weber, S. Withanalage, the team at the MPI for Geoanthropology / Science of Human History, and all Glyph players. The research leading to this work was funded by the Max Planck Society, PSL university, and the Agence Nationale de la Recherche.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2025.104620.

Data availability

Open data and code can be found at this address: https://osf.io/ewp68/.

References

Antic, G., & Altmann, G. (2005). On letter distinctivity. *Glottometrics*.

Bruner, A., & Kasdan, M. L. (2001). Handwriting errors: Harmful, wasteful and preventable. *The Journal of the Kentucky Medical Association, 99*(5), 189–192.

Chandra, S., Bokil, P., & Udaya Kumar, D. (2015). Anatomy of Bengali Letterforms: A Semiotic Study. In A. Chakrabarti (Ed.), *ICoRD'15 – Research into Design Across Boundaries Volume 1* (pp. 237–247). Springer India. https://doi.org/10.1007/978-81-322-2232-3_22.

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B: Biological Sciences, 272*(1560), 267–275. https://doi.org/10.1098/rspb.2004.2942

Clements, G. N. (2003). Feature economy in sound systems. *Phonology, 20*(3), 287–333.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition, 163*, 128–145. https://doi.org/10.1016/j.cognition.2017.02.001

Dehaene, S. (2010). *Reading in the Brain: The New Science of How We Read* (Reprint edition). Penguin Books.

Dunbar, E., & Dupoux, E. (2016). Geometric Constraints on Human Speech Sound Inventories. *Frontiers in Psychology, 7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01061.

Engesser, S., & Townsend, S. W. (2019). Combinatoriality in the vocal systems of nonhuman animals. *WIREs Cognitive Science, 10*(4), e1493.

Fernandes, T., & Leite, I. (2017). Mirrors are hard to break: A critical review and behavioral evidence on mirror-image processing in developmental dyslexia. *Journal of Experimental Child Psychology, 159*, 66–82. https://doi.org/10.1016/j.jecp.2017.02.003

Galantucci, B., Kroos, C., & Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies, 11*(1), 100–111. https://doi.org/10.1075/is.11.1.03gal

Roberts, G., & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and Cognition, 4*(4), 297–318. https://doi.org/10.1515/langcog-2012-0017

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences, 12*(10), 381–387. https://doi.org/10.1016/j.tics.2008.06.006

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2023). *Glottolog 4.8*. Doi: 10.5281/zenodo.8131084.

Han, S. J., Kelly, P., Winters, J., & Kemp, C. (2022). Simplification is not dominant in the evolution of chinese characters. *Open Mind, 6*, 264–279. https://doi.org/10.1162/opmi_a_00064

Hockett, C. F. (1960). The origin of speech. *Scientific American, 203*, 4–12.

Hockett, C. F. (1966). *The quantification of functional load—A linguistic problem*. https://eric.ed.gov/?id=ED011649.

Jee, H., Tamariz, M., & Shillcock, R. (2022). Systematicity in language and the fast and slow creation of writing systems: understanding two types of non-arbitrary relations between orthographic characters and their canonical pronunciation. *Cognition, 226*, Article 105197. https://doi.org/10.1016/j.cognition.2022.105197

Kelly, P., Winters, J., Miton, H., & Morin, O. (2021). The predictable evolution of letter shapes: an emergent Script of West Africa recapitulates historical change in writing systems. *Current Anthropology, 62*(6), 669–691. https://doi.org/10.1086/717779

King, A. (2018). The lexicon is optimised for a noisy channel. *Glottometrics, 43*, 58–67.

Kirby, S., & Tamariz, M. (2021). Cumulative cultural evolution, population structure and the origin of combinatoriality in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences, 377*(1843), Article 20200319. https://doi.org/10.1098/rstb.2020.0319

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Köhler, R. (1987). System theoretical linguistics. *Theoretical linguistics, 14*(2–3), 241–258. https://doi.org/10.1515/thli.1987.14.2-3.241

Lachmann, T., & Geyer, T. (2003). Letter reversals in dyslexia: Is the case really closed? A critical review and conclusions. *Psychologische Beitrage, 45*(Suppl1), 50–70.

Ladd, D. R. (2014). On duality of patterning. In D. R. Ladd (Ed.), *Simultaneous Structure in Phonology* (p. 0). Oxford University Press. Doi: 10.1093/acprof:oso/9780199670970.003.0005.

Ladefoged, P. (2000). *Vowels and Consonants: An Introduction to the Sounds of Languages*. Wiley-Blackwell.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lally, C., & Rastle, K. (2023). Orthographic and feature-level contributions to letter identification. *Quarterly Journal of Experimental Psychology*, *76*(5), 1111–1119. Doi: 10.1177/17470218221106155.

Levy, R. (2008). *A noisy-channel model of rational human sentence comprehension under uncertain input*. EMNLP. In *08: Proceedings of the conference on Empirical Methods in Natural Language Processing* (pp. 234–243). https://doi.org/10.3115/1613715.1613749

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*(4), 839–862. https://doi.org/10.2307/411991

Little, H., Eryılmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition, 168*, 1–15. https://doi.org/10.1016/j.cognition.2017.06.011

Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution, 20*(3), 116–121. https://doi.org/10.1016/j.tree.2004.12.002

Mackie, S., & Mielke, J.. (2011). Feature economy in natural, random, and synthetic inventories. In Ridouane, & G. N. Clements (Eds.), *Where Do Phonological Features Come From?*.

Marcet, A., & Perea, M. (2017). Is nevtral NEUTRAL? Visual similarity effects in the early phases of written-word recognition. *Psychonomic Bulletin & Review, 24*(4), 1180–1185. https://doi.org/10.3758/s13423-016-1180-9

Martinet, A., 1971. Langue et fonction. Dénoël Gonthier.

Meletis, D. (2020). *The Nature of Writing: A Theory of Grapholinguistics*. Fluxus editions.

Mielke, J. (2008). *The Emergence Of Distinctive Features*. Oxford University Press.

Miton, H., & Morin, O. (2021). Graphic complexity in writing systems. *Cognition, 214*, Article 104771. https://doi.org/10.1016/j.cognition.2021.104771

Morin, O. (2018). Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy. *Cognitive Science, 42*(2), 664–677. https://doi.org/10.1111/cogs.12550

Morin, O., & Miton, H. (2018). Detecting wholesale copying in cultural evolution. *Evolution and Human Behavior, 39*(4), 392–401. https://doi.org/10.1016/j.evolhumbehav.2018.03.004

Nowak, M. A., Krakauer, D. C., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society of London B: Biological Sciences, 266*(1433), 2131–2136. https://doi.org/10.1098/rspb.1999.0898

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46*, 4646–4674.

Perea, M., Abu Mallouh, R., Mohammed, A., Khalifa, B., & Carreiras, M. (2018). Does visual letter similarity modulate masked form priming in young readers of Arabic? *Journal of Experimental Child Psychology, 169*, 110–117. https://doi.org/10.1016/j.jecp.2017.12.004

Pouplier, M. (2020). Articulatory phonology. *In M. Aronoff (Ed.), Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.745

Primus, B. (2004). A featural analysis of the modern roman alphabet. *Written Language & Literacy, 7*(2), 235–274. https://doi.org/10.1075/wll.7.2.06pri

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition, 141*, 52–66. https://doi.org/10.1016/j.cognition.2015.04.001

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2021). *A language of thought for the mental representation of geometric shapes*. PsyArXiv. Doi: 10.31234/osf.io/28mg4.

Sandler, W. (2008). The Syllable in Sign Language: Considering the Other Natural Language Modality. In B. Davis & C. Zajdo (Eds.), *Ontogeny and Phylogeny of Syllable Organization, Festschrift in Honor of Peter MacNeilage*. Taylor Francis.

Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural Language & Linguistic Theory, 29*(2), 503–543. https://doi.org/10.1007/s11049-011-9128-2

Sanguineti, V., Laboissière, R., & Payan, Y. (1997). A control model of human tongue movements in speech. *Biological Cybernetics, 77*(1), 11–22. https://doi.org/10.1007/s004220050362

Sarró, R. (2023). *Inventing an African Alphabet: Writing, Art, and Kongo Culture in the DRC*. Cambridge University Press.

Scott-Phillips, T. C., & Blythe, R. A. (2013). Why is combinatorial communication rare in the natural world, and why is language an exception to this trend? *Journal of The Royal Society Interface, 10*(88), Article 20130520. https://doi.org/10.1098/rsif.2013.0520

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sohn, H.-M. (2001). *The Korean Language*. Cambridge University Press.

Tamariz, M., & Kirby, S. (2015). Culture: Copying, Compression, and Conventionality. *Cognitive Science, 39*(1), 171–183. https://doi.org/10.1111/cogs.12144

Therneau, T., Atkinson, B., port, B. R. (producer of the initial R., & maintainer 1999-2017). (2023). *rpart: Recursive Partitioning and Regression Trees* (Version 4.1.23) [Computer software]. https://cran.r-project.org/web/packages/rpart/index.html.

van der Hulst, H., & van der Kooij, E. (2020). Sign language phonology: Theoretical perspectives. In J. Pfer, R. Pfau, & A. Herrmann (Eds.), *Routledge Handbook of Theoretical and Experimental Sign Language Research*. Routledge.

Vaux, B., & Samuels, B. (2015). Explaining vowel systems: Dispersion theory vs natural selection. *The Linguistic Review, 32*(3), 573–599. https://doi.org/10.1515/tlr-2014-0028

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics, 43*, 57–68. https://doi.org/10.1016/j.wocn.2014.02.005

Verhoef, T., Kirby, S., & de Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive Science, 40*(8), 1969–1994. https://doi.org/10.1111/cogs.12326

Wiley, R. W., Wilson, C., & Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology. Human Perception and Performance, 42*(8), 1186–1203. https://doi.org/10.1037/xhp0000213

Zuidema, W., & de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics, 37*(2), 125–144. https://doi.org/10.1016/j.wocn.2008.10.003

Zuidema, W., & de Boer, B. (2018). The evolution of combinatorial structure in language. *Current Opinion in Behavioral Sciences, 21*, 138–144. https://doi.org/10.1016/j.cobeha.2018.04.011