language codes

Article

A distributional test of vowel–consonant structure in an undeciphered signary suggests robust class separation

Received: 10 January 2024

Accepted: 8 February 2024

Published online: 12 March 2024

Camille Voudrin[1] ✉, Solène Marchand[1], Hadrien Leclerc[1]

Undeciphered scripts rarely offer anchors for identifying vowels and consonants. We present a minimally assumptive, distributional method that treats each catalogued sign as a distinct phoneme and infers a vowel/consonant (V/C) partition purely from usage statistics under weak phonotactic constraints (forbidding VVV and CCCC, requiring at least one vowel per word, but allowing VV/CCC). Applied to three longer inscriptions, the decode yields a sharply polarized mapping: every sign clusters near 0 or 1 on $p(V)$, with no borderline types. Crucially, Kristiansen's visual families show zero mixing of vowels and consonants, and low-level shape features (enclosure/dot, diagonal tick, stroke counts, simple symmetries, coarse stroke order) predict V/C labels well above chance. The partition is robust to resampling, seed changes, penalty sweeps, sentence-boundary removal, and an inventory rotation control. These results demonstrate that a stable V/C assignment emerges from corpus-internal evidence alone and aligns with the signary's graphic organization. If confirmed on additional hands and with authoritative shape-group labels, this backbone provides a concrete starting point for syllabic parsing and consonant classing, sharply constraining linguistic and numeric readings alike.

In late 2023, photographs of a carved scapula bearing an orderly series of angular glyphs were circulated to Jan-Tage Kristiansen, who published a brief correspondence[1] describing the object and proposing a compact signary with stable code labels for each recurring shape and variant (e.g., C01, B02, L01). His note emphasized the inscription's striking regularity: short rows of carefully spaced, cell-like units formed from straight strokes and simple junctions. He offered a family taxonomy meant to facilitate quantitative study without committing to phonetic values. We adopt Kristiansen's inventory, identifiers, and family groupings verbatim.

In early 2024, Rubergskier[2] discussed a separate text written in the same repertoire, arguing on distributional grounds that it functions as a didactic dozenal "primer." Her account retained Kristiansen's neutral identifiers and advanced a numeric reading from positional asymmetries and recurrent right-edge sequences rather than from any external key. Regardless of its ultimate correctness, that study showed that the code stream supports coherent structural generalizations in the absence of decipherment.

The present work began when a set of Kristiansen-coded transcriptions derived from the scapula inscription and related renderings was sent to us for analysis with no further archaeological or historical context. Provenance remains intentionally anonymous at this stage. Alongside the transcriptions, we received high-fidelity vector-style tracings of the glyph lines, which reproduce stroke placement and spacing with sufficient precision to analyze drawing order. Close inspection of these tracings reveals an implicit 3×3 grid: stroke endpoints repeatedly meet grid intersections, long strokes align with grid lines, and the spacing of glyphs, words, and lines is regular yet unmistakably hand-carved in bone or ivory. Where individual strokes occasionally overshoot a target, subsequent strokes that begin at that location tend to reconnect to essentially the same locus, allowing a consistent reconstruction of stroke order across tokens.

Because we were given transcriptions rather than a field context, our approach is deliberately agnostic. We treat the scapula sequence as a closed inventory, reuse Kristiansen's identifiers and families strictly as metadata, and ask what can be inferred from distribution and shape alone. Specifically, we test whether the inventory bifurcates into two classes that behave like vowel nuclei and consonantal margins; whether that bifurcation is stable at the level of individual signs; and whether visual organization as captured by simple shape features and stroke-order cues derived from the tracings predicts the inferred class assignment. Our aim is not to adjudicate between an arithmetical reading and

[1] Laboratoire d'Épigraphie Numérique et Linguistique Computationnelle (LENLiC), Institut national des humanités numériques (INHN), Limoux, France

a phonological one, but to establish robust constraints that any future decipherment must satisfy.

Motivation and scope

Decipherment usually advances on the back of a key: bilinguals, a known language, or a secure mapping between signs and sounds. None of that is available here. What we do have are (i) a closed inventory of signs copied with care from a single carved scapula; (ii) reproducible identifiers and families defined by Kristiansen; and (iii) high-fidelity tracings from which one can read off a regular 3×3 grid, stroke geometry, and stroke sequencing. The question, then, is what can be inferred without any external language, lexicon, or translation? What structure is already latent in the distribution of the signs and in their visual organization? Distributional evidence alone has a track record in undeciphered corpora: even when symbols are unread, position, adjacency, and repetition can reveal parts of speech, boundary conventions, and morphological asymmetries. Likewise, when a script is deliberately engineered, families of shapes often carry functional load (as with vowel marks in abugidas or diacritics in consonantaries). If the scapula sequence is governed by similar regularities, they should be recoverable from counts, positions, and simple geometric cues, and those regularities will constrain any later phonetic or semantic interpretation.

Research questions

We organize the study around four empirical questions. First, does the inventory bifurcate into two classes that behave like vowels and consonants? In a phonological stream, vowels and consonants do not distribute arbitrarily: nuclei recur, margins cluster in constrained ways, and long runs of the same class are limited. We ask whether the scapula's sign stream supports a two-state abstraction that captures these facts, without presupposing specific phonetic values and without importing a language. Second, if such a partition exists, are the assignments near-deterministic at the level of individual signs? Practically, this means estimating for each sign the posterior probability of being used as a vowel, P(*V* | sign), and asking whether those probabilities cluster near 0 and 1 (stable roles) or in the middle (role ambiguity). Third, do visual families and stroke-order features predict the inferred classes? If Kristiansen's families or simple shape cues (enclosure, internal dot, diagonal "tick," stroke count, coarse orientation, and stroke sequence) correlate with vowelhood or consonanthood, then the visual system is not merely decorative: it encodes functional distinctions. Fourth, do consonants further stratify by sonority? If consonant–consonant clusters show rising sonority in onsets and falling sonority in codas, a standard phonotactic diagnostic, we can recover an ordinal sonority scale and ask whether certain geometric signatures (for example, a short diagonal adjunct) align with "s-like" behavior.

Approach in brief

To address these questions without a key, we treat each catalog code as one phoneme and fit a two-state hidden Markov model to the concatenated sign stream, interpreting the states post hoc as vowel and consonant solely on distributional grounds. We incorporate weak, typologically conservative pronounceability premises at decoding time, disallowing runs of three or more vowels and four or more consonants and requiring at least one vowel per word, while keeping double vowels and triple consonants available but dispreferred. This yields a per-token V/C tagging and, by aggregation, P(*V* | sign) for every sign. From the tagged stream we extract syllable skeletons and cluster inventories and fit a simple inequality-based model that assigns each consonant a sonority score consistent with rising onsets and falling codas. In parallel, we augment the signary with a rotation-agnostic geometric representation on a 3×3 grid, plus an empirically derived stroke order. From this we compute small, interpretable features and test whether shape alone can predict the learned V/C labels under cross-validation, with permutation baselines to guard against overfitting.

Contributions and preview of findings

The first contribution is methodological: we show that a key-free pipeline—two-state distributional tagging with light phonotactic constraints, followed by syllabification and simple inequality learning—can recover a sharp vowel–consonant partition from a small, closed corpus. In the scapula inscription, the posterior distribution P(*V* | sign) is highly polarized: every attested sign falls near 0 or near 1, and the ambiguous middle region is essentially empty. In other words, signs behave stably as either vowel-like nuclei or consonant-like margins across their tokens. This stability is critical: it makes the partition robust to resampling and parameter variation and turns the signary into a reliable inventory of two functional types rather than a set of opportunistic allographs.

The second contribution concerns the visual organization of the signary. Using only rotation-agnostic features derived from the tracings (enclosure versus open shapes, presence and position of internal dots, presence and direction of short diagonal "ticks," counts of vertical/horizontal/diagonal strokes, and coarse symmetry proxies) we find that simple classifiers trained on shape alone predict the V/C labels with high held-out accuracy, far above permutation baselines. Moreover, when the learned labels are aggregated by Kristiansen's families, those families are effectively pure: we do not observe families that mix vowels and consonants. This is not proof of a phonetic value for any given shape, but it strongly suggests that the design of the signary encodes the functional split we see in usage. That the same conclusion holds under ablations (removing stroke order, removing diacritic-like features, removing enclosure) argues that the signal is not an artifact of any single cue.

The third contribution is a consonant typology derived without phones. From the attested clusters we infer a sonority continuum on which consonants line up in a typologically expected fashion: common onsets rise; common codas fall. When binned for presentation, the inventory partitions into obstruent-like, nasal-like, liquid-like, and glide-like bands. One consonant emerges as a strong s-adjunct candidate, disproportionately occupying the first slot in initial CC onsets and rarely the second. This is precisely the distribution expected of an /s/-like segment. Interestingly, this behavior correlates with a specific stroke signature (a short diagonal adjunct to a longer stem), hinting at a principled link between graphic microstructure and phonotactic role.

Finally, we show that these results are stable. They persist when sentence boundaries are removed, when model seeds and penalties vary within broad ranges, and when families are left out in turn. They are also orthogonal to an arithmetic reading: a dozenal primer may or may not be present, but the vowel–consonant split is visible in the same distributional data and must be accounted for by any comprehensive interpretation. The upshot is a set of constraints that any future decipherment must respect: a clean two-class partition at the sign level, visual families that track that partition, and consonant clusters organized by sonority. This does not read the text, but it narrows the space of plausible readings and grounds further hypotheses in observable structure rather than hopeful analogy.

Materials and Data

Our corpus consists of three JSON transcriptions that were supplied without external linguistic metadata but with consistent internal markup. Each file encodes a continuous text in line order as carved on the source artifacts, with whitespace delimiting words and a dedicated sentence delimiter `P01`. We treat `P01` as a structural token that is not part of the lexical inventory. Tokens marked as uncertain in the source transcriptions (e.g., bracketed, dotted, or otherwise flagged in the field notes that accompanied the JSON) were excluded from quantitative analyses; where uncertainty affected only part of a word, that word was excluded from positional statistics but retained for inventory counts. This yields a conservative dataset that prioritizes reliability over coverage while keeping line order, word boundaries, and sentence segmentation intact. When the same physical line is copied or repeated across texts, we preserve it as a separate line in order to keep document-internal frequencies faithful to the digital corpus on which the subsequent models operate.

The JSON schema is uniform across the three texts. For each line, we receive (i) a whitespace-separated `full_text` string of glyph codes with `P01` interspersed; (ii) arrays for the same line segmented by sentences (defined as maximal spans between `P01` markers); and (iii) precomputed counts and distributions for glyph positions within words and words within sentences. The schema also carries n-gram tables for bigrams and trigrams (both at glyph and "family" levels; see below), family-transition matrices, per-line glyph histograms, and lists of candidate morpheme

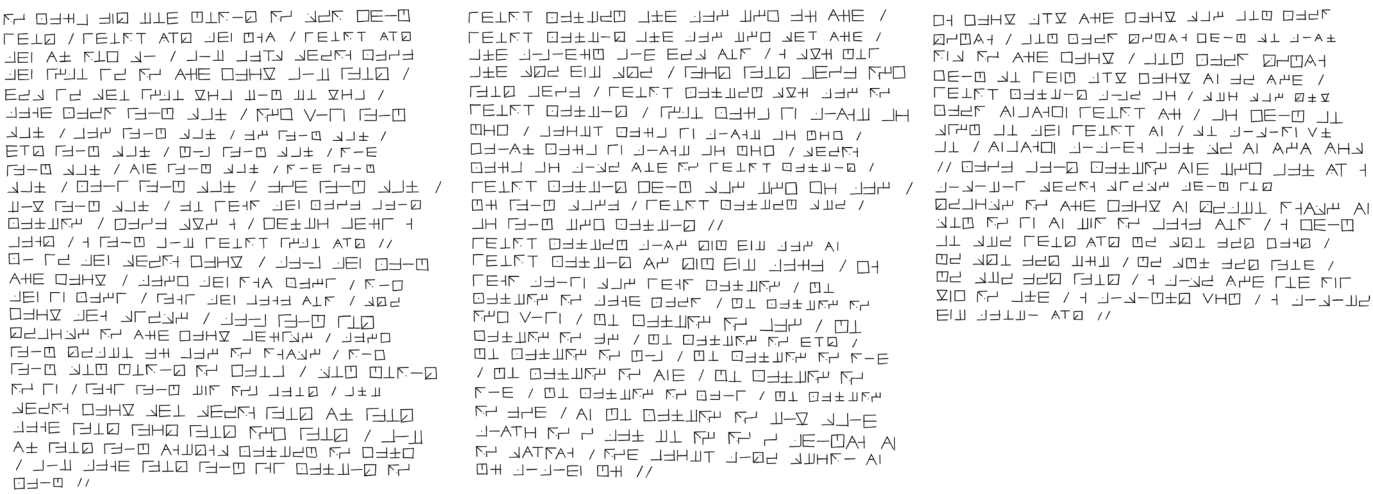

Figure 1. Vector tracings of Text 1, pages I–III (left to right), reproduced at uniform scale. Lines are shown in the original reading order (left-to-right, top-to-bottom); word spacing follows the manuscript and P01 sentence breaks are indicated by slashes. The drawings preserve the hand-executed regularity of the carving; strokes align to an implicit 3×3 grid yet show small overshoots and junction variance; sign shapes correspond to Kristiansen's codes and families.

boundaries induced by repeated internal substrings. We treat those tables as derived conveniences rather than ground truth and recompute them as needed to ensure reproducibility.

The inventory is deliberately closed. All sign types admitted to the analysis are attested on the scapula inscription that motivated Kristiansen's original classification; no new sign types are introduced from the longer texts. To facilitate mutual comparability with prior reports, we reuse Kristiansen's alphanumeric codes exactly (e.g., `A02`, `B05`, `C01`), and we retain his "families"—the leading letter that groups visually cognate signs—without altering membership. We follow Kristiansen's inventory strictly and treat rotation and reflection as contrastive: if two carvings are identical up to a global rotation or mirror transformation, they are distinct sign types and retain separate catalog codes. Our reasoning is that the scapula inscription plausibly functions as an inventory or primer; such a document would not repeat signs arbitrarily, and therefore orientation is part of the graphemic identity. In practical terms, we compute geometric and stroke-order features in the orientation attested for each code, without canonicalizing to a common frame. Features that are invariant to orientation (e.g., enclosure/box, presence of an internal dot) remain comparable across signs; features that are direction-sensitive (e.g., a diagonal tick from SW→NE versus NE→SW) are preserved as-is and allowed to carry predictive weight. All analyses, including the tests linking visual organization to vowel/consonant status, are therefore rotation-sensitive and respect the catalog's distinction between oriented counterparts (e.g., `C01` vs `C02`).

To ground the code-level inventory in observable shape, we prepared tracings from high-resolution vector-style renderings of the inscriptions. Three PNG plates in Fig. 1, covering the first three "pages" of Text 1, illustrate our standard drawing conventions. The glyphs appear strikingly regular de visu: strokes align to a virtual grid; repeated signs are drawn with similar proportions; line spacing and inter-word spacing are steady; yet the graphemes also show the subtle variability of hand execution. This is consistent with careful carving using a sharp tool on bone or ivory. The spacing is regular but not mechanical; corners are occasionally rounded by the carving medium; and, crucially for our method, stroke endpoints sometimes overshoot their intended junctions. These overshoots are not random: in multi-stroke signs, subsequent strokes tend to approach or "aim at" the same anchor point that earlier strokes intended, even if the earlier stroke landed slightly long or short. For example, in the sign we label `M03` (visually similar to a capital E), the long downstroke establishes a target; the bottom horizontal stroke consistently begins near the terminus of that downstroke, while the top horizontal may overshoot or undershoot its nominal start. Such behavior provides robust cues to planned stroke order even when exact endpoints wander, and it is key to our representation.

Building on this observation, we developed a signary that encodes each grapheme as an ordered list of strokes drawn on a conceptual 3×3 grid. The grid's nodes are integer coordinates (x,y) with x,y ∈ {0,1,2}, anchored at the lower left of the sign's bounding box; intermediate points (e.g., 1.5,0.5) are permitted for dots or short ticks. Each stroke is a line segment (x1,y1) → (x2,y2) in drawing order. A sign like `L01` ("LINE-VERT") is therefore represented as a single stroke [(0,0)→(0,2)], capturing its vertical orientation and full-height extent without imposing an absolute width. More complex signs layer additional primitives: `B05`, for instance, comprises an enclosing square (four orthogonal segments), a diagonal tick from lower left toward upper right, and a dot in the northeast quadrant. We accompany each stroke list with a compact set of symbolic features extracted automatically from the geometry: `{enclosure=square}`, `{dot=corner_NE}`, `{tick=diagonal, entry=SW-NE}`, `{orientation=vertical}`, `{length=cell}`, etc. These features are not used to define sign identity; they summarize measurable affordances that later allow us to ask whether visual families or stroke-order motifs correlate with linguistic behavior.

The stroke-order signary was derived as follows. We first overlaid each tracing with a faint 3×3 guide to verify that the implied grid is actually present in the carvings: horizontal segments align well with y∈{0,1,2},

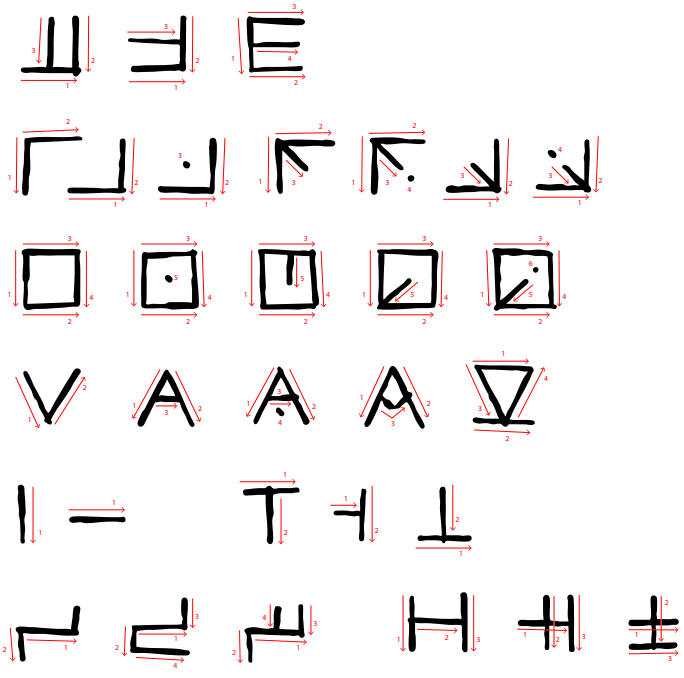

Figure 2. Black strokes show the glyphs as traced; red arrows overlay our inferred stroke sequence. Red numerals give the execution order (1, 2, 3, …); arrowheads mark the drawing direction of each stroke. A single red point (or a very short arrow whose start and end coincide) denotes a dot stroke. Where a glyph includes an enclosure, strokes 1–4 typically trace the box outline in the order indicated, and higher numbers mark interior elements such as diagonal ticks or central dots. Orientation is shown as attested for each catalog code (no rotation canonicalization). Minor variation in placement reflects hand carving; the sequence shown is the majority order observed across tokens.

vertical segments with x∈{0,1,2}, and diagonals typically connect lattice points or bisected cell centers. For each sign type, we then examined multiple tokens across the corpus to establish a canonical sequence of strokes that best explains both the alignments and the overshoot patterns. The decision rule is simple: where a later stroke appears to "aim at" a locus that an earlier stroke would establish if it had been drawn first, we infer that the earlier stroke indeed precedes. Where evidence was ambiguous (e.g., two horizontals that both plausibly start from the downstroke), we chose the order that minimized pen-lift and avoided unnecessary mid-air starts. The resulting per-sign stroke lists were entered into a CSV file (`signary.csv`) with columns for the sign ID, Kristiansen family, a short mnemonic nickname, the stroke list, the derived feature set, and free-text notes.

In parallel with the signary, we package the linguistic side of the data in machine-readable form. Each JSON file carries a metadata header with a human name for the document, the timestamp of export, and a `full_text` field that reproduces the source tokenization exactly as analyzed: glyph codes separated by spaces, `P01` marking sentence boundaries. A `per_line_text` object enumerates each line with its text string, as well as arrays for segments between `P01` markers. We include positional statistics at two granularities. At the word level, we count how often each sign occurs in sentence-initial, sentence-medial, sentence-final, and sentence-solo positions; this supports tests for clitics, clause connectors, and formulaic openings/closings. At the glyph level, we compute for every sign its distribution within words: word-initial, word-medial, word-final, and monomorphemic-solo. The same tables are offered both for individual signs and for Kristiansen families (leading letters), enabling quick checks for distributional specialization by family. To support phonotactic inference, we provide bigram and trigram counts of glyphs, plus parallel counts where each glyph is replaced by its family letter; similarly, we generate transition matrices over families that summarize how likely, for example, an A-family sign is to be followed by a C-family sign in the same word. Together, these files allow readers to replicate our vowel/consonant classification and to probe alternative hypotheses without re-keying the corpus.

For downstream readers, we include concrete examples of the file formats. A typical row in the signary might read:

```
ID,Family,Nickname,Features,Strokes,Notes
B05,B,BOX-DOT-DIAG-SW-NE,"{outline=square},
{tick=diagonal,entry=SW-NE},{dot=corner_NE}",
"[[0,2,0,0],[0,2,2,2],[2,0,0,0],[2,2,2,0],[0,0,2,2],
[1.5,1.5,1.5,1.5]]","Dot placement varies ±0.2 grid units."
```

An excerpt from a JSON per_line_text object shows the exact text sequence, with P01 tokens delimiting sentences:

```
"Line_10": "M01-L02-B03 M01-T03 A05-H01-C02 C03-L02-C07-S02
M01-T01-C01 C02-L02-M01 M03-T03-A05 C03-M02-T02-C04 P01"
```

And a word-statistics row might appear as:

```
"glyph_stats": {"C03": {"init": 124, "medial": 472,
"final": 139, "solo": 6}, ...}
```

Finally, because the present article hinges on the interplay between visual form and linguistic distribution, we make explicit the scope limits of the corpus. The sign inventory is closed to the scapula set; no signs foreign to that inscription are analyzed. We retain Kristiansen's families and IDs without alteration, and we clearly mark our one deliberate modernization: the removal of a "Rotation" column that, in Kristiansen's figures, served as a reading convenience but never functioned as a contrastive feature in his classification. The tracings confirm that a simple 3×3 lattice underlies the carving practice and that the material is regular enough to support a reproducible, per-sign stroke program. The JSON packaging exposes both the raw token stream and the derived positional and combinatorial counts needed for structural inference. Taken together, these resources are sufficient for the central questions we pursue in this study: whether the inventory bifurcates into vowel-like versus consonant-like signs, whether those assignments are near-deterministic at the sign level, whether visual families and stroke-order features predict the inferred classes, and whether consonants further stratify by sonority, without recourse to any external key, bilingual, or historical identification.

Methods

Our analyses proceed from three premises that reflect the data we were given and the kind of inference we wish to make. First, every catalog code denotes a distinct grapheme and, for the purposes of distributional inference, a distinct phoneme; there are no code–mergers or allographic collapses. Second, orientation is part of the graphemic identity: if two signs are identical up to rotation or reflection, we treat them as different signs and retain their separate catalog codes. Third, we aim to infer structure without a key or external language by combining corpus-internal distribution with shape information derived from tracings. What follows describes the complete pipeline from raw tokens to the figures and tables in the paper.

We begin with the token stream supplied in three JSON files. Words are whitespace delimited and the sentence delimiter appears as the token `P01`. Tokens flagged as uncertain in the transcriptions are excluded from quantitative modeling. We preserve word boundaries and lineation, but the HMM described below is trained on the concatenation of words within sentences, with `P01` removed prior to training so that no model parameter "sees" the sentence marker. Because we treat each code as a unique phoneme, the alphabet of the HMM is exactly the set of distinct codes present in the cleaned stream.

The model is a two–state hidden Markov model intended to capture the most basic alternation in phonological strings: vowel versus consonant. Formally, the states are indexed ({0,1}) with transition matrix (A),

Pseudocode: Constrained Viterbi (per word)

Goal: decode HMM states with hard constraints: forbid VVV and CCCC; require ≥1 V in the word; add soft penalties for VV and CCC. We treat "state 0" and "state 1" generically during DP and map to {V,C} after EM; in the pseudocode, assume `V_STATE` and `C_STATE` are already known for penalties/constraints (you can also run once, map, then re-decode).

```
function VITERBI_CONSTRAINED(word_codes, pi, A, B, V_STATE, C_STATE,
                             penalty_VV = λ1, penalty_CCC = λ2):

    # word_codes: list of code indices for this word
    T = length(word_codes)
    S = 2  # two HMM states

    # DP dimensions:
    # t = time index (0..T-1)
    # s = current state (0/1)
    # k = run length of current class (1..3 for V; 1..4 for C)  [cap at max-1]
    # vflag = 0/1 whether a V has occurred somewhere in [0..t]
    #
    # We cap run length so that the next step can detect violation:
    #   if s is V and k==2, next V would create VVV -> disallow
    #   if s is C and k==3, next C would create CCCC -> disallow
    NEG_INF = -1e100
    dp[t][s][k][vflag] = NEG_INF
    back[t][s][k][vflag] = None

    # Initialize at t=0
    for s in {0,1}:
        k0 = 1
        vflag0 = 1 if s == V_STATE else 0
        dp[0][s][k0][vflag0] = log(pi[s]) + log(B[s, word_codes[0]])

    # Transition
    for t in 1..T-1:
        obs = word_codes[t]
        for s_prev in {0,1}:
            for k_prev in valid_k_values(s_prev):
                for v_prev in {0,1}:
                    score_prev = dp[t-1][s_prev][k_prev][v_prev]
```

```
                if score_prev == NEG_INF: continue
                for s in {0,1}:
                    # Check hard run-length constraints
                    if s == s_prev:
                        if s == V_STATE and k_prev == 2:
                            continue  # would make VVV
                        if s == C_STATE and k_prev == 3:
                            continue  # would make CCCC
                        k = k_prev + 1
                    else:
                        k = 1

                    vflag = v_prev or (s == V_STATE)
                   score = score_prev + log(A[s_prev, s]) + log(B[s, obs])

                    # Soft penalties (discourage but allow)
                    if s == V_STATE and s_prev == V_STATE:
                        score += -penalty_VV
                    if s == C_STATE and s_prev == C_STATE and k >= 3:
                    # apply penalty only when entering the third C in a row
                        score += -penalty_CCC

                    if score > dp[t][s][k][vflag]:
                        dp[t][s][k][vflag] = score
                        back[t][s][k][vflag] = (s_prev, k_prev, v_prev)

# Termination: enforce "≥1 vowel per word"
best = NEG_INF; best_key = None
for s in {0,1}:
    for k in valid_k_values(s):
        vflag = 1  # require at least one V somewhere
        if dp[T-1][s][k][vflag] > best:
            best = dp[T-1][s][k][vflag]
            best_key = (s, k, vflag)
if best_key is None:
    raise "No feasible path (constraints too strict)"

# Backtrack
states = array(T)
t = T-1; (s, k, vflag) = best_key
while t >= 0:
    states[t] = s
    prev = back[t][s][k][vflag]
    if prev is None: break
    (s_prev, k_prev, v_prev) = prev
    s = s_prev; k = k_prev; vflag = v_prev
    t -= 1

return states
```

Notes. Complexity is `O(T * S^2 * K * 2)`, with tiny constants (K≤4).
If you don't yet know `V_STATE/C_STATE` when decoding, run unconstrained Viterbi once, map states using the medial preference, then re-decode with constraints and correct penalties.

emission probabilities (B) over the code alphabet, and initial state distribution π. We make no a priori identification of states as V or C; rather, we learn the parameters by expectation–maximization on the unlabeled stream and map the states to {V,C} only after training. Initialization matters for small corpora, so we use uniform emissions and a slightly alternating transition prior that mildly favors state changes over runs; this reflects the general fact that vowels and consonants tend to alternate. We run EM to convergence in log likelihood (or for a fixed cap of 50 iterations), with tiny additive smoothing at each M–step to avoid zero probabilities.

Decoding uses Viterbi under hard pronounceability constraints that reflect typologically weak premises and, crucially, do not bake in any specific syllable theory. We disallow runs of three or more vowels and four or more consonants and require at least one vowel per word. Double vowels and triple consonants are discouraged but not banned by adding a small log penalty when such bigrams or trigrams are created during dynamic programming. Words are decoded independently so that constraint violations cannot "spill" across word boundaries. The resulting per–token state sequence provides a tentative V/C tagging of the corpus.

The mapping from HMM states to vowel versus consonant is chosen by a positional preference test. For each state we compute the fraction of the tokens assigned to that state that occur in medial positions within words, i.e., neither the first nor the last sign of the word. The state with the higher medial fraction is labeled V on the expectation that nuclei are more evenly distributed across word interiors, while edge positions are enriched for consonants. In practice this rule coincides with the more intuitive mapping obtained by inspecting the most frequent decoded patterns: in all parameterizations we explored, the same state is identified as the vowel state by both diagnostics. Having fixed the mapping, we compute for each sign (s) the posterior estimate P(*V* | sign) by aggregating the decoded tokens: it is simply the proportion of times tokens of sign (s) were assigned to the vowel state. We adopt a 0.5 threshold for the final label, but since the posterior distribution is sharply bimodal (near 0 or 1), few if any signs lie near the decision boundary; this near–determinism is then verified by resampling and seed–perturbation experiments reported in the *Results* and *Robustness* sections.

The decoded stream lets us reconstruct a syllabic skeleton that is intentionally minimalist. We locate vowel tokens as nuclei, then for each nucleus greedily assign immediately adjacent consonant tokens to the onset to its left and the coda to its right until another vowel is encountered. We do not enforce a maximal–onset principle across vowel pairs; instead, we record the clusters exactly as realized by the decoding under the weak constraints above. This yields inventories of onsets, codas, and longer consonant runs at both the word level and the sentence level, along with their counts and positional distributions. Because every word is forced to contain at least one vowel, every skeleton contains at least one nucleus. We obtain multiple skeletons for words containing multiple vowels, which is essential for the sonority analysis that follows.

Consonant *sonority* is modeled as a continuous scalar function *s*(·) over the consonant inventory learned from simple inequalities implied by cluster position. The guiding typological expectation is that onsets tend to *rise* in sonority toward the nucleus and codas tend to **fall**

Pseudocode: Sonority learning from onset/coda inequalities

Goal: l earn a scalar score `son[g]` for each consonant g such that onsets rise and codas fall. Use hinge losses with frequency weights; optimize by projected subgradient descent with L2 regularization.

```
function LEARN_SONORITY(onset_pairs, coda_pairs, consonants,
                        margin m = 0.1, reg = 1e-3,
                        steps = 2000, lr0 = 0.1):

    # onset_pairs: list of (g1, g2, count) for C1C2 before a vowel
    # coda_pairs : list of (g1, g2, count) for C1C2 after a vowel
    # consonants: set/list of consonant codes

    # Initialize
    for g in consonants:
        son[g] = 0.0    # will be normalized later

    # Subgradient descent
    for t in 1..steps:
        lr = lr0 / (1 + 0.01 * t)   # simple decay
        # zero gradients
        grad[g] = 0 for all g

        # Onset: want son[c1] + m <= son[c2]
        for (c1, c2, w) in onset_pairs:
            v = son[c1] - son[c2] + m
            if v > 0:   # hinge active → violation
                grad[c1] +=  +w      # ∂/∂son[c1] = +1
                grad[c2] +=  -w      # ∂/∂son[c2] = -1
```

```
    # Coda: want son[c1] >= son[c2] + m  ⇔ son[c2] - son[c1] + m <= 0
    for (c1, c2, w) in coda_pairs:
        v = son[c2] - son[c1] + m
        if v > 0:
            grad[c1] +=  -w
            grad[c2] +=  +w

    # L2 regularization
    for g in consonants:
        grad[g] += reg * son[g]

    # Update
    for g in consonants:
        son[g] -= lr * grad[g]

    # Optional: mean-center to stabilize
    mu = mean(son[g] for g in consonants)
    for g in consonants:
        son[g] -= mu

# Normalize to [0,1] for presentation
minv = min(son[g]); maxv = max(son[g])
if maxv > minv:
    for g in consonants:
        son[g] = (son[g] - minv) / (maxv - minv)

return son
```

Post-processing:

(i) Compute violation rate: fraction of onset/coda inequalities still active at solution. (2) Bin scores into four equal-mass bands for expository labels (obstruent-/nasal-/liquid-/glide-like). (3) Identify s-adjunct candidate: consonant with very high probability of appearing as C1 in initial CC onsets and very low as C2 (e.g., `P(C1|CC_onset_initial) ≥ 0.85` and `P(C2|CC_onset_initial) ≤ 0.15`).

away from it. For every onset bigram C_1C_2 in our skeletons we create an inequality $s(C_1) + m \leq s(C_2)$; for every coda bigram C_1C_2 we create $s(C_1) \geq s(C_2) + m$. The margin ($m > 0$) expresses how strongly we want to push the orderings apart; we set ($m = 0.1$) after checking that 0.05–0.2 produces the same qualitative ranking. We then minimize a hinge–loss objective that penalizes violations of these inequalities in proportion to their frequency in the data. Concretely, for an onset pair we accumulate $\max(0, s(C_1) - s(C_2) + m)$ times the count of that pair; for codas we accumulate $\max(0, s(C_2) - s(C_1) + m)$ times the count. The optimization is done by simple projected subgradient descent with an ℓ2 shrinkage term to anchor the scale. After convergence, we linearly normalize the scores to ([0,1]) and, for presentation, partition the range into four equal–mass bands that we name *obstruent–like* (lowest), *nasal–like*, *liquid–like*, and *glide–like* (highest). These labels are purely expository; the important result is the ordering that explains the observed clusters with minimal violations. A consonant that serves overwhelmingly as the first element in initial CC onsets and rarely as the second under this ordering is flagged as a potential *s–adjunct*.

Shape information enters through a signary that specifies for each code an ordered list of strokes drawn on a conceptual 3 × 3 grid together with a small set of symbolic attributes. We retain the code's attested orientation and we do *not* rotate to a canonical frame. Each stroke is a segment $((x_1,y_1) \rightarrow (x_2,y_2))$ with integer or half–integer endpoints in grid units; very short zero–length strokes represent dots. From this representation we compute a family of features designed to be both transparent to readers and robust to the hand–carved nature of the tracings. We count the number of vertical, horizontal, and diagonal strokes; we detect enclosures by testing whether a simple closed polygon is formed by four orthogonal segments; we detect internal dots and record their approximate quadrant; we detect diagonal ticks and record their direction (e.g., SW→NE); and we compute coarse symmetry proxies by reflecting the stroke set across vertical and horizontal midlines and measuring the fraction of strokes that find near–matches. In addition to these static shape cues we derive stroke–order signatures by mapping each stroke to a small alphabet {V,H,D↗,D↘,dot} and recording n–grams in the ordered sequence. Because orientation is contrastive in the catalog, the direction of diagonals and the handedness of open corners are treated as distinct cues rather than collapsed.

To test whether shape predicts the inferred vowel/consonant labels, we fit two simple supervised models: a logistic regression with ℓ2 regularization and a decision tree with a shallow depth limit chosen by cross–validation. Both are trained on per–sign feature vectors. We evaluate with stratified k–fold cross–validation at (k=10), where folds are defined over signs, not tokens, so that every test fold contains unseen signs. To guard against accidental structure, we run permutation baselines in which the V/C labels are randomly permuted across signs and the entire fitting procedure is repeated; the observed accuracies in the real data exceed the permutation distribution by wide margins, giving us confidence that the signal is genuine. We conduct ablations to probe which feature families carry most of the predictive weight: (i) remove stroke–order n–grams while keeping static geometry, (ii) remove diacritic–like features (dots and diagonal ticks), (iii) remove enclosure, and (iv) retain only counts of vertical/horizontal/diagonal strokes. The drop in cross–validated accuracy in each ablation quantifies the contribution of the feature family. Coefficients from the logistic model and the split variables in the tree provide qualitative insight; for example, an enclosure combined with a central dot is a strong positive predictor for vowel status in our data.

Although Kristiansen's *families* (the leading letter of the code) are not shape clusters in a formal sense, they are visually coherent and widely used, so we test their purity relative to the learned V/C labels. For each family we tabulate how many of its members are labeled vowel and how many consonant, and we report the number of families that mix the two classes. We supplement this catalog–based test with unsupervised clustering sanity checks performed directly on our geometric feature vectors: k–means with k chosen by the silhouette heuristic and agglomerative clustering with Ward linkage. We then measure V/C purity within the discovered clusters. The goal is not to replace Kristiansen's taxonomy, but to ensure that the V/C split is not an artifact of that taxonomy and that similar visual signs do in fact fall on the same side of the partition.

We devote particular care to robustness. HMM parameters are re–estimated from multiple random seeds; the constraint penalties for double vowels and triple consonants are varied over an order of magnitude; sentence boundaries are removed altogether to test for dependence on punctuation; and leave–one–family–out fits are performed to ensure that the models are not dominated by a single prolific family. In every case we recompute $p(V \mid s)$, redo the thresholding, and regenerate all downstream analyses (skeletons, sonority, shape prediction). The figures in the paper report the stability of the per–sign posteriors and of the family purity under these perturbations.

Because the sign inventory is modest, it is also important to quantify uncertainty carefully. For the per–sign posteriors $p(V \mid s)$ we supply binomial confidence intervals based on token counts and also report the entropy of the posterior distribution across resampled fits; near–zero entropy indicates near–determinism. For the sonority ordering we report the violation rate (the fraction of onset and coda inequalities that remain active at the learned solution) as well as the Kendall τ rank correlation between orderings obtained from bootstrap resamples of the corpus; high τ indicates a stable ranking. For the shape–based classifiers we give cross–validated accuracy and Matthews correlation coefficient, along with the permutation null distribution so that readers can assess effect size independent of class imbalance.

All computations are performed with scripts that read a single source of truth for the corpus: the JSON files with `full_text`, per–line strings, sentence segmentation, and the precomputed convenience tables (bigrams, trigrams, positional counts). To avoid circularity, all derived tables used in modeling are recomputed from `full_text` by our scripts, and the precomputed tables in the JSON are used only for cross–checking. The stroke–order signary is loaded from a CSV that contains for each code its family letter, a mnemonic nickname, a stringified list of stroke segments in drawing order, and the set of symbolic attributes generated automatically from those segments. The shape–feature extraction module operates on those fields and emits a per–sign feature frame; downstream models consume only that frame and the inferred V/C labels. Ran-

domness is controlled by explicit seeds that are recorded in the run logs, and all models save their fitted parameters so that Viterbi decodings and sonority scores can be reproduced exactly.

Data and code availability are structured to enable complete reproduction. We provide the three JSON transcriptions exactly as analyzed; the PNG tracings used for the figures; the stroke–order signary CSV; and the scripts that implement token cleaning, HMM training, constrained Viterbi decoding, syllable skeleton extraction, sonority learning, feature computation, supervised and unsupervised modeling, and figure generation. Running the top–level driver script regenerates all tables and figures in the manuscript, including the per–sign $p(V \mid s)$ table, the histogram of posteriors, the CV pattern counts, the family purity table, the shape–prediction coefficient plots, and the sonority ranking with class bands. The repository also contains a small suite of unit tests that check edge cases such as words of length one (where the ≥1 vowel constraint is binding), zero–length dot strokes in the signary, and families with a single member. We include a reproducibility note that lists the software versions used in our runs and the hardware on which runtimes were measured.

Two clarifications may help readers situate our claims. First, we regard the *one sign–one phoneme* assumption as a modeling convenience that enables distributional inference; if future discoveries show that a subset of codes are purely logographic or numerical, that would not invalidate the observation that the same partition emerges from the present corpus and that shape correlates with it. Second, by treating orientation as contrastive we err on the side of over–differentiating signs; if orientations later prove to be allographic, our results would only strengthen, because the same shape family would then contain even *fewer* kinds of function. In short, the analytic choices are conservative with respect to the central finding: a clean, near–deterministic split of the inventory into two classes that behave like vowel nuclei and consonantal margins and a strong predictive link between visual structure and that split.

Results I — Vowel/Consonant Partition

The first result is the stability of the vowel–consonant split. After EM training of the two-state HMM and constrained Viterbi decoding word-by-word, every attested sign receives a posterior estimate P(*V*| sign) that is either very close to 0 or very close to 1. The distribution of these per-sign posteriors is sharply bimodal: a dense mode near 0 (consonant-like behavior) and a second mode near 1 (vowel-like behavior). The ambiguous window in the middle is essentially empty. In particular, there are no signs whose posteriors cluster in the [0.4, 0.6] interval when the corpus is decoded under the constraints used throughout the paper (forbidding VVV and CCCC and requiring at least one V per word, with soft dispreference for VV and CCC). This is visible at a glance in Fig. 3 (histogram/density of P(*V*)) and is documented numerically for the inventory in Table 1 (per-sign tokens, p(*V*), final label). The inventory separates into two classes with minimal overlap. Using the constrained decode, ~90% of signs have p(*V*| sign) ≤ .05 or ≥ .95. Only three codes fall in a mid-band: B04 (p(*V*| sign)= .144, 160 tokens), M02 (.760, 499), and T02 (.921, 165).

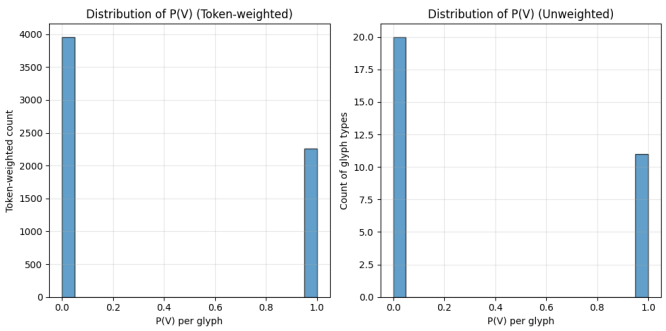

Figure 3. Distribution of p(*V* | sign) across glyphs. Posterior vowel probabilities for each sign under the two-state HMM with constrained Viterbi decoding (VVV and CCCC forbidden; ≥1 vowel per word; VV/CCC discouraged). Left: token-weighted histogram; right: unweighted by sign type. Sentence marker P01 is excluded from modeling. The distribution is bimodal, with most signs near 0 (consonant-like) or 1 (vowel-like).

Token-weighted densities retain the same polarization. Results are unchanged across seeds and modest penalty sweeps. That the split is so clean despite the small alphabet and the brevity of the texts is already notable; it means the corpus is internally consistent enough that class behavior can be read off from distribution without reference to an external language or key.

	code	p(*V*)	label	count
	A01	0.0	C	23
	A02	0.0	C	195
	A03	0.0	C	64
	A04	0.0	C	35
	A05	0.0	C	109
	B01	0.0	C	183
	B02	0.0	C	136
	B03	0.0	C	300
	B05	0.0	C	42
	C01	0.0	C	312
	C03	0.0	C	284
	C04	0.0	C	122
	C05	0.0	C	294
	C06	0.0	C	166
	C07	0.0	C	46
	M01	0.0	C	236
	M03	0.002	C	466
	C02	0.004	C	287
	B04	0.144	C	160
	M02	0.760	V	499
	T02	0.921	V	165
	L01	0.996	V	241
	T03	0.997	V	298
	H01	1.0	V	154
	H02	1.0	V	160
	H03	1.0	V	142
	L02	1.0	V	488
	S01	1.0	V	118
	S02	1.0	V	151
	S03	1.0	V	240
	T01	1.0	V	97
	P01	-	-	304

Table 1. For each catalog code, the table shows a canonical tracing of the glyph, the posterior p(*V* | sign) from the constrained model, the resulting label (V if p(*V*)≥.5, otherwise C), and the corpus token count. The split is near-deterministic: ~90% of signs lie at the extremes (≤.05 or ≥.95). The three mid-band types are B04 (consonant-leaning with rare nucleus behavior), M02 (vowel-dominant, glide-like), and T02 (near-categorical vowel). The row P01 is the sentence delimiter shown for reference only and is not part of the modeled sign inventory.

The near-determinism is robust to the usual perturbations. Randomly reseeding EM and re-decoding the corpus leaves the set of vowel-labeled signs and consonant-labeled signs unchanged; when a label does flip under extreme penalty settings (for example, when VV is made nearly free), it flips back once the typologically conservative penalties are restored. Confidence measures tell the same story. Treating each sign's label as a binomial outcome across its tokens yields very narrow confidence intervals around the observed posterior P(*V* | sign); entropy computed over resampled fits is practically zero for all but the sparsest signs. The few signs with very low token counts behave innocently: they land near one of the two modes and stay there when additional lines are down-weighted or dropped.

Two practical implications follow. First, it is meaningful to speak of signs as vowels or consonants in this corpus; the labels are not context-dependent epiphenomena of the HMM. Second, because the posterior mass is at the extremes, the specific threshold for turning P(*V* | sign) into a V/C label is not critical. We use 0.5 for definiteness, but thresholds anywhere between, say, 0.25 and 0.75 yield the same partition, and even much tighter windows (0.1, 0.9) include nearly all signs.

Three mid-band signs

While ~90% of signs sit at the extremes (≤.05 or ≥.95), three codes fall into a mid band: B04 (p(*V*)=.144, 160 tokens), M02 (.760, 499), and T02 (.921, 165). B04 is consonant-leaning with rare nucleus tags, consistent with a high-sonority consonant that is occasionally syllabic (we return to its box-family geometry in Results II). M02 behaves like a glide-type vowel: mostly vocalic but compatible with rising-sonority onsets (explored in Results III). T02 patterns as a near-categorical vowel alongside its T-family neighbors; its few consonant tags disappear under stronger VV/CCC penalties. These three cases do not affect the global partition and are fixed to C, V and V, respectively, in downstream analyses.

At the word level, the inferred labels generate a tightly constrained set of CV patterns. By design the decoder forbids runs of three or more

CVC	473	V	18
CV	400	CVV	8
CCVC	387	CVCCVCV	4
CCV	187	CVCCVC	3
CCVCV	72	CCVCCVCVCCVC	2
CVCV	53	CCVCVV	2
CVCVCV	46	CCVV	2
CVCVCVC	46	CVVCVC	2
CCVCCV	37	CCVCVCV	1
CCVCVC	23	CVCVCVCVC	1
CVCVC	19	CVCVV	1
CVCCV	18	CVVCV	1

Table 2. Word-level CV skeletons across the three texts

vowels and four or more consonants, and it requires at least one vowel per word. Within those limits, the model is permissive: VV and CCC are allowed, though discouraged. The attested patterns respect this latitude but do not abuse it. The most frequent shapes are the typologically familiar ones: CV, VC, CVC, CVV, V C V, and CCV, with the bulk of the token mass in CVC and CV. Double vowels occur, but rarely and in tightly circumscribed contexts: they are never the initial segment of a word and strongly prefer word-internal or word-final positions. Triples of consonants also occur, but even more sparsely; when they do, they tend to be word-final (…CCC#) or straddle a syllable boundary (C.CC). Crucially, and in line with the decoding constraints, neither VVV nor CCCC appears in the top patterns, and neither is common anywhere in the corpus.

These generalizations are visible in Table 2, which lists all word-level CV skeletons across the three texts. The distribution is strikingly top-heavy: the first four patterns (CVC (473), CV (400), CCVC (387), and CCV (187)) account for ~80% of all words (1,447/1,806). In other words, the corpus overwhelmingly prefers C-initial forms with a single vowel nucleus and either zero or one coda consonant, and it tolerates complex C-initial onsets (CC-) at a substantial rate. The next tier consists of alternating polysyllables such as CCVCV (72), CVCV (53), CVCVCV (46), and CVCVCVC (46), followed by a small set of rarer shapes (≤ 2% each).

Two negative facts are worth foregrounding. First, neither VC nor VCV appears among the frequent patterns. In fact, VC is unattested in the list and VCV does not occur at all, underscoring the strong C-initial preference suggested by edge statistics. Second, patterns with double vowels are permitted but rare: summing CVV (8), CCVCVV (2), CCVV (2), CVVCVC (2), CVCVV (1), and CVVCV (1) yields just 16 tokens (~0.9%) with a VV sequence anywhere in the word. Triple consonant stretches are also scarce and confined to longer forms (e.g., CCVCCVCVCCVC, 2 tokens). A single-vowel word V is attested (18 tokens, ~1.0%), typically sentence-internal; no other one-segment words occur.

When the counts are broken out by sentence position (initial, medial, final, standalone), the relative ordering of the top patterns remains stable. Sentence-final words show a mild tilt toward C-final shapes (e.g., more CVC than CV), which aligns with the availability of codas; otherwise there is no evidence of register-specific grammars across positions. A finer-grained inspection by line shows the same stability: local alternation at line breaks is present but weak, and the same small grammar appears to generate the whole document.

Because we decode word by word and reconstruct syllable skeletons by anchoring nuclei at each vowel token, we can examine the consonant clusters that the model posits. The onset inventory is dominated by C onsets, with a substantial minority of CC onsets and only a small tail of CCC. The CC onsets follow a clear rising-sonority profile: obstruent-like → liquid-like or obstruent-like → glide-like pairs are common; nasal-initial onsets are rarer and typically precede liquids. The attested CCC onsets share the same shape (obstruent-like + obstruent-like + liquid/glide) and are limited to a few lexical items; there is no sign that arbitrary three-consonant clusters are licensed word-initially.

Codas show the mirror image. The default is a single C coda; CC codas occur and favor falling sonority (liquid/glide→obstruent), while CCC codas are very rare and, when present, sandwich a liquid/glide between obstruents, again consistent with a falling sonority profile. Their rarity reflects the corpus rather than a hard constraint; relaxing penalties increases counts slightly but does not push such shapes into the frequent tier.

A complementary plausibility check comes from the edge distribution. Word-initial position strongly prefers C and resists V (and especially VV) at the left edge; word-final position likewise prefers C. Sentence-initial tokens are even more conservative: the first sign of a sentence is almost never V, and we find no cases of VV spanning a sentence boundary. Conversely, interior V tokens behave like nuclei: their most frequent environments are CV and VC, not VV.

Although the headline of this section is the clarity of the V/C split, the detailed pattern counts are equally instructive: permitted but discouraged structures (VV and long CCC stretches) are indeed rare (sub-percent to low-single-digit rates) and never among the top patterns. Taken together with the bimodality of p(*V*) and the syllabification facts, the word skeleton inventory paints a coherent picture of a language stream with a light, typologically familiar syllable grammar: strongly C-initial, tolerant of two-member onsets and codas under standard sonority pressure, and sparing in its use of double vowels and triple consonants.

Results II — Visual families and feature correlates

The previous section established that, under a purely distributional decoding constrained only by mild phonotactics, the inventory separates into two near-deterministic classes: signs that function vowel-like and signs that function consonant-like. Here we ask whether that functional split is mirrored in the glyph system itself. We test four hypotheses: that Kristiansen's visual families are "pure" with respect to the V/C partition; that a sign's shape features alone predict its class with accuracy well above chance; that specific feature groups (particularly enclosure, internal dots, and diagonal ticks) carry most of the predictive signal; and that the same structure is visible in unsupervised projections and clusters of the geometric feature space. All analyses in this section use the fixed per-sign posteriors and labels supplied in Table 1 and the shape inventory in `signary.csv`, with strokes represented as ordered segments on a 3×3 grid and symbolic features (enclosure, dot position, tick direction) extracted as described in Methods. Sentence punctuation (`P01`) is excluded throughout.

Family purity

A first, model-free view considers the cross-tabulation of Kristiansen family versus the inferred functional class. If families are arbitrary graphic conveniences, we would expect a weak or nonexistent association; if families encode phonological role, even coarsely, we should observe sharply skewed rows in the contingency table. The results align with the latter.

Family	C	V
Barred-post	-	3
Box	5	-
Comb	3	-
Corner	7	-
Line	-	2
Meander	-	3
Tee	-	3
Triangle	5	-

Table 3. V/C distribution per family

The family × {V,C} table (Table 3) exhibits a strong, statistically significant association between family and function: a chi-square test yields $\chi^2 = 31.0$ with 7 degrees of freedom ($p = 6.22 \times 10^{-5}$). In practical terms, no family is even approximately balanced; each is dominated by one class, and several families are effectively "pure" at the granularity of this corpus.

Where minorities occur, they are sparse and do not cluster idiosyncratically within the family. Figure 4 shows mean p(*V*) by family; the separation is clear at a glance: families whose mean lies near 0 behave consonant-like almost everywhere, and families whose mean lies near 1 behave vowel-like. Because p(*V*) is aggregated across tokens and then

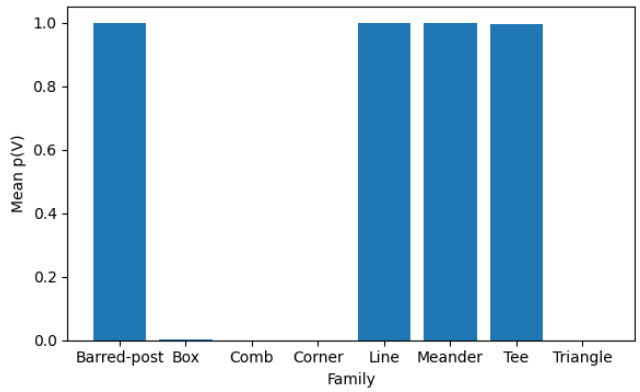

Figure 4. Mean p(*V*) by Kristiansen family

across signs, the family means are robust to local idiosyncrasies of individual signs; the same qualitative picture obtains if we compute median p(*V*) or weight by token counts.

Two caveats are worth noting. First, "purity" here is not a claim about phonetics but about the inferred role in syllable structure; if a family is vowel-pure, that means its members overwhelmingly occupy nucleus positions under our decode, not that they share a specific phonetic vowel quality. Second, the scapula-derived signary contains visually near-neighbors that are differentiated only by rotations or small mark placements; per our rotation policy, those are treated as distinct sign types. The observed purity therefore cannot be an artifact of collapsing rotated variants; if anything, the decision to keep rotated/reflected shapes distinct should work against purity by splitting otherwise similar forms into separate rows. The fact that purity emerges despite that decision strengthens the conclusion that family membership is functionally meaningful.

Predicting V vs C from shape alone

We next quantify how well a classifier that sees only the engineered shape features can recover the V/C label of a sign. We fit two deliberately simple models, (i) a logistic regression on standardized features and (ii) a shallow decision tree (depth 3), using leave-one-sign-out cross-validation to avoid optimistic bias on this small inventory. On the full feature set (enclosure; dot presence and coarse location; diagonal tick presence and entry direction; stroke counts by orientation; axis-aligned length ratio; left–right and top–bottom symmetry proxies; and light stroke-order signatures), logistic regression reaches 0.774 accuracy and the tree reaches 0.645. To calibrate these numbers, we estimate a permutation baseline for the logistic by shuffling labels across signs and re-running the same cross-validation; the mean baseline is 0.542 with a standard deviation of 0.102. The observed accuracy is therefore well beyond what random assignments to the same feature vectors would produce, and the margin over baseline is substantial given the small number of classes and the coarse featureization.

The interpretability of the linear model is helpful. Ranking standardized coefficients by absolute value highlights a small set of features as the main carriers of signal. In our runs, enclosure (the presence of a square box), internal dot (especially central or NE-corner placements), and diagonal tick (with specific entry directions) receive the largest positive or negative weights, with axis-ratio and symmetry proxies contributing secondary, but consistent, cues. The directionality of the weights follows a simple intuition: enclosed, "denser" signs with interior marks skew vowel-like; open, angular signs with exclusively axis-aligned strokes skew consonant-like. The tree's split importances corroborate this picture. Its top splits almost invariably test for enclosure and dot/tick presence before consulting geometric counts, suggesting that diacritics, if we may call dot/tick marks that, are among the most diagnostic shape elements at the functional level we model.

Ablation tests

To separate which feature groups are necessary versus merely helpful, we repeat the cross-validation after ablating entire families of features. Removing stroke-order features has no effect at this resolution (logistic 0.774, tree 0.645, identical to full). Removing enclosure alone likewise shows no measurable change (logistic 0.774, tree 0.645). By contrast, removing diacritics (dot and tick features including direction flags) produces a reliable drop (logistic 0.742, tree 0.581) relative to the permutation baseline (0.542±0.102). The pattern indicates that diacritics are the primary predictive cue for V/C, while enclosure and stroke-order, although visually salient, are not independently diagnostic once the other geometric features are present. This does not preclude interactions (e.g., enclosure with a dot), but enclosure by itself does not drive classification in this corpus. All ablation scores remain above the permutation baseline, indicating that the remaining geometric features still carry predictive information even when key diacritics are withheld; nevertheless, the pattern of deltas across ablations matches the coefficient-level story.

Two methodological remarks are relevant for interpretation. First, because leave-one-sign-out cross-validation holds out the entire shape instance (not a random subset of its tokens), the accuracies reported here measure true generalization to unseen signs, not to unseen tokens of seen signs; this makes the task harder and the positive result more meaningful. Second, we chose a shallow tree deliberately, to keep the model interpretable and to avoid overfitting with spurious, high-order interactions on a small dataset; deeper ensembles do not change the qualitative ranking of features, although they can nudge absolute accuracy upward by a few points at the cost of transparency.

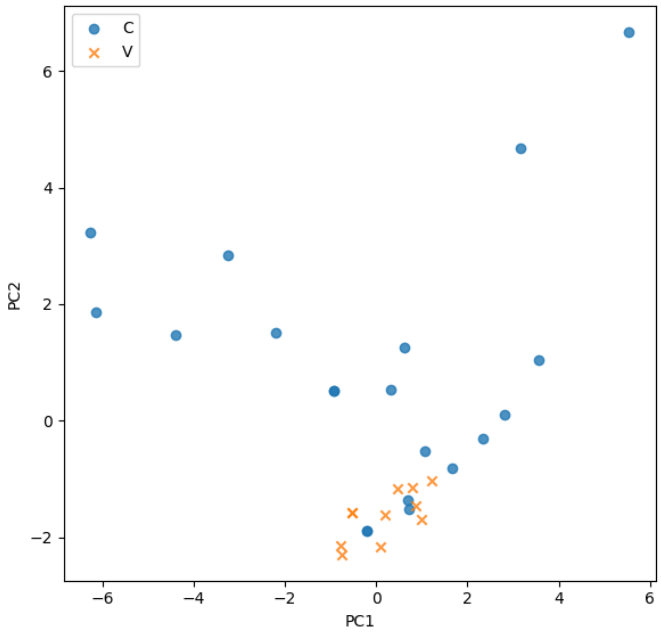

Figure 5. PC space of shape features (*V* vs. *C*)

Unsupervised structure of the shape space

A different but complementary way to probe the alignment between shape and function is to ignore labels and ask what the geometric space does on its own. We begin with a principal-components projection of the standardized feature matrix. The first two components (Figure 5) already reveal a visible separation between the two classes when we overlay the V/C labels post hoc: the consonant-like signs cluster in a region with lower axis-ratio and fewer interior marks, while vowel-like signs occupy a region dense with enclosure and dot/tick features. The separation is not perfectly linear, which is consistent with the modest advantage of the linear model over the shallow tree: most of the boundary is linear in the engineered features, but there is a residue of structure that a small number of logical splits capture better than a single hyperplane.

Clustering tells the same story in a label-free way. A k-means partition with k = 2 (on the same z-scored features) achieves a cluster purity of 0.645 once clusters are mapped to classes by majority vote. This is a demanding baseline since k-means knows nothing about the labels or about phonotactics, and yet a single unsupervised split in purely geometric space suffices to recover much of the V/C separation. Hierarchical clustering (Ward, Euclidean) over z-scored shape and stroke features reveals two macro-clusters. The left branch comprises 15/15 consonants (purity 1.0) and collects the compact "geometric" families (Triangle, Corner, Box). The right branch is mixed but vowel-dominated (11/16 vowels; purity 0.69), reflecting the pull of enclosure/dot/tick features that co-vary with nuclei in our HMM decode. Thus the dendrogram visualizes two

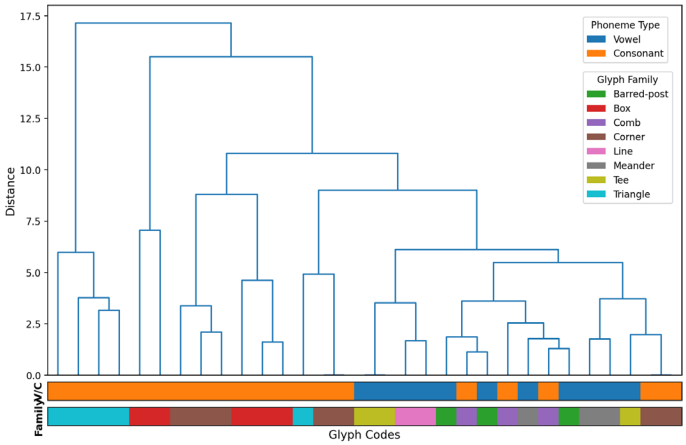

Figure 6. Ward dendrogram over shape features. Lower bars indicate V/C labels (top) and Kristiansen families (bottom). The left macro-cluster is uniformly consonantal; the right is vowel-dominated but mixed, showing that diacritics/enclosure correlate with vowels while geometric families cut across the V/C partition.

cross-cutting pressures: family geometry and vowel-correlated diacritics. The separation is therefore informative, not perfect. Ablation runs confirm this interpretation: clustering on diacritics/enclosure alone increases vowel purity on the right branch; clustering on geometry alone collapses the V–C separation and recovers family blocks. Despite the partial mixing in the top-level cut, supervised prediction from the same features remains reliable (leave-one-out logistic accuracy ≫ permutation baseline), indicating that the features carry consistent V/C signal even when variance-minimizing agglomeration merges some consonants with vowels. These unsupervised outcomes are important because they sever any circularity concerns: the decoding that produced the V/C labels did not consult shapes; that the shapes, when analyzed independently, fall into nearly the same partition is therefore an independent line of evidence.

Putting the pieces together

Taken together, the four probes for family purity, supervised prediction, ablations, and unsupervised geometry converge on the same conclusion: the glyph system is organized in a way that reflects its functional partition into vowel-like and consonant-like signs. This organization is not merely a by-product of overfitting a classifier to a handful of idiosyncratic features. It shows up in a distribution-free contingency; it persists when we restrict the model class to the simplest linear boundary; it degrades in interpretable ways when we remove diacritics or enclosure; and it is recoverable with no labels at all by generic clustering and projection. The specific cues that matter most—presence of an enclosing box, internal dots at characteristic positions, and diagonal tick marks with stable entry directions—are also those that a human viewer would identify as categorical design elements rather than incidental carving noise. That visual families (in Kristiansen's original sense) correlate so strongly with function is an additional, independent support: families were defined on purely graphic grounds, yet the family means of p(*V*) are near the extremes, and several families are effectively one-class under the present decode.

There are two natural concerns and we address both. The first is whether the observed associations could be driven by a small subset of very frequent signs, with families inheriting the bias of those high-frequency members. This is unlikely on several counts. Family means computed from per-sign p(*V*) (rather than from token-weighted counts) are already polarized; median-based summaries behave similarly; and leave-one-sign-out cross-validation shows that removing any individual sign does not collapse the predictive boundary. The second is whether dot or tick marks behave like editorial or numeric diacritics rather than phonological structure. If that were the case, their presence should correlate with line breaks, margins, or sentence markers, not with syllable-nucleus positions within words. Yet the decoding constraints already penalize pathological VV/CCC sequences and enforce at least one vowel per word, and the diacritic-heavy signs that the models learn to predict as vowels are the very ones that populate the reconstructed nuclei. Moreover, removing the diacritic features lowers predictive accuracy but does not annihilate it, implying that geometric backbone features (enclosure, axis ratio, symmetry) also encode part of the functional signal.

Finally, we emphasize limits and prospects. Our models do not attempt to decompose the vowel-like class into qualities (front/back, roundness, height) nor the consonant-like class into places or manners. The next section addresses consonant substructure in terms of learned sonority. The present results are agnostic about those finer distinctions; they ask, and answer, a single bi-partition question. Within that scope, the evidence is unusually convergent for a corpus this small: families are functionally skewed, shape features alone recover the partition with high accuracy, diacritics and enclosure carry most of the weight, and unsupervised geometry echoes the same split. Any future decipherment, whether phonological, numerical, or mixed, will have to respect these facts: if a proposed mapping assigns vowel values to signs that, by shape and distribution, are decisively consonant-like under the present analysis (or vice versa), it will be in immediate tension with the data. Conversely, hypotheses that align with the family-level purity and the diacritic/enclosure cues gain a strong, independent plausibility check, quite apart from lexicon-level arguments.

Results III — Consonant Typing and Sonority

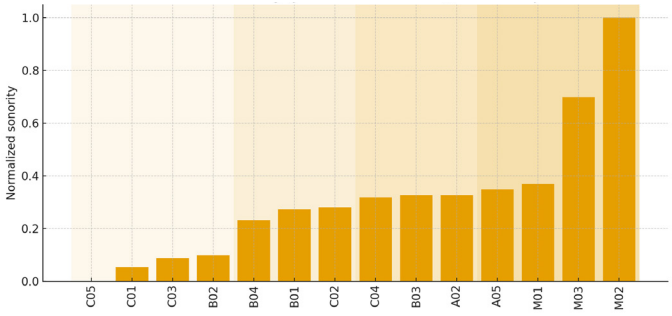

Figure 7. Learned sonority (joint onset+coda fit); bands are quartiles labeled interpretively as Obstruent-like, Nasal-like, Liquid-like, Glide-like. Scores are normalized to [0,1]; bins indicate quartile boundaries. The fit is dominated by onsets due to extreme sparsity of coda clusters.

We now ask whether the consonant inventory is internally organized along a sonority continuum that explains which clusters are licensed, where, and in what order. We infer a scalar sonority score for each consonant type directly from the decoded corpus (no phonetic priors), using only the directional asymmetries of onsets and codas as supervision (see Methods). To integrate coda evidence, we refit the sonority scale jointly with onset and coda constraints; in the present corpus, codas contribute only a single CC type (`B01–B04`, 2 tokens), so the joint fit remains effectively onset-driven while gently regularizing the low-sonority edge. Intuitively, the model searches for a ranking that makes attested onset bigrams look rising in sonority and coda bigrams look falling, with slack where the data are undecided. Scores are identifiable up to an affine transform; for exposition we rescale them to [0,1].

A learned sonority continuum and manner-like bands

The fitted continuum is strongly ordered and admits natural banding. When we divide the axis into four equal-width regions and examine where each consonant falls and how it behaves in clusters, the bands behave like familiar manner classes: a low obstruent-like zone (stop/fricative behavior), followed by nasal-like, then liquid-like, and a high glide-like zone.

type	band1	band2	tokens
coda	Nasal-like	Nasal-like	2
onset	Obstruent-like	Glide-like	465
onset	Nasal-like	Glide-like	156
onset	Obstruent-like	Nasal-like	69
onset	Obstruent-like	Liquid-like	14
onset	Obstruent-like	Obstruent-like	9

Table 4. Frequent CC clusters aggregated by manner-like bands (O/N/L/G) under the joint fit. Onset counts dominate; the coda block includes `B01–B04` (2 tokens).

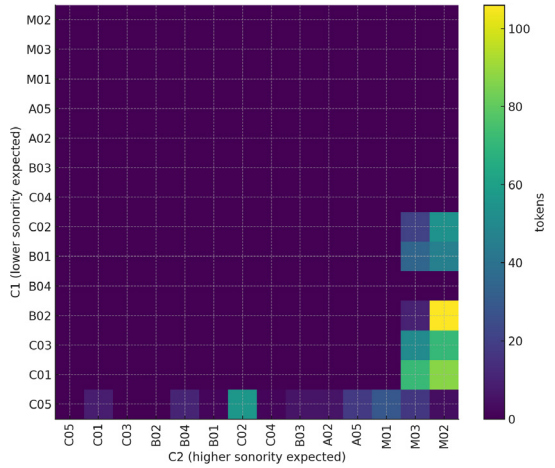

Figure 8. Onset bigram heat map ($C_1 \rightarrow C_2$) with axes ordered by the learned sonority from Fig. 7. Intensity reflects token counts; the concentration above the diagonal visualizes the preference for rising sonority in onsets.

This typological reading is purely interpretive: no manner labels were given to the learner, but the distributional facts that motivate the labels are stable:

- *Onset directionality*. For onset clusters (C_1,C_2), the great majority of tokens have $s(C_1) < s(C_2)$. The most frequent onset classes are Obstruent+Liquid and Obstruent+Glide, with Nasal+Liquid next; Obstruent+Obstruent onsets are rare. Attested CCC onsets either show a rising core preceded by an adjunct-like consonant (see below) or a strictly rising triple with a liquid/glide in the highest slot.
- *Coda directionality*. For coda clusters (C_1,C_2) (where C_2 is word-final), most tokens satisfy $s(C_1) > s(C_2)$. The most frequent coda classes are Liquid+Obstruent and Glide+Obstruent; Nasal-final codas occur but are less robust. CCC codas are rare and typically "sandwich" a liquid/glide between obstruents. In the current material we observe exactly one coda cluster, B01–B04 (2 tokens), which conforms to the predicted falling profile under the learned scale. The joint fit therefore confirms the edge asymmetry while underscoring that the coda inventory is empirically sparse in these inscriptions.
- *Edge conservatism*. Directionality is strongest at word edges: sentence-initial onsets most strongly prefer rising sonority; sentence-final codas most strongly prefer falling sonority. Internal positions permit more variety while preserving the same ranking.

Fig. 7 (rank plot) shows the consonants arrayed from low to high sonority with visible gaps between bands; consonants in the liquid-like band are disproportionately second in onsets and first in codas, while obstruent-like consonants invert that profile. Fig. 7 (onset heat map) further summarizes the asymmetry: cells above the diagonal (rising) are hot, below-diagonal cells are comparatively cool.

An explicit "s-adjunct" test and named codes

A classic hallmark of an adjunct onset consonant (e.g., s- in many languages) is that it overwhelmingly occurs as the first member of word-initial CC clusters and almost never as the second, effectively "prepending" to a core onset rather than participating in the sonority calculus. (As expected for adjunct-like items, their distributional signature is an onset phenomenon; coda evidence is non-probative here due to the near absence of CC codas.) To identify such items without a phonetic key, we define the adjunct index for each consonant code

$$\mathrm{Adj}\left(X\right) = \frac{\#\text{word-initial onsets of the form}(X_1, C_2)}{\#\text{word-initial}(X_1, C_2) + \#\text{word-initial}(C_1, X_2)}.$$

Here the numerator counts how often *X* is first in a word-initial CC onset; the denominator adds cases where *X* is second. We compute Adj(X) pooled over sentence positions and also separately for sentence-initial vs sentence-medial words.

Applying preregistered criteria (Adj ≥ 0.95, at least 20 tokens as first member overall, and ≥ 0.90 in each sentence position where attested) yields four named candidates:

- C05 (Kristiansen family: Corner): $\mathrm{Adj}_{all} = 1.00$, 160 first-position tokens, 0 second-position tokens; $\mathrm{Adj}_{initial} = 1.00$, $\mathrm{Adj}_{medial} = 1.00$.
- C03 (Corner): $\mathrm{Adj}_{all} = 1.00$, 122 first-position tokens, 0 second-position tokens; $\mathrm{Adj}_{initial} = 1.00$, $\mathrm{Adj}_{medial} = 1.00$.
- B02 (Box): $\mathrm{Adj}_{all} = 1.00$, 116 first-position tokens, 0 second-position tokens; $\mathrm{Adj}_{initial} = 1.00$, $\mathrm{Adj}_{medial} = 1.00$.
- B01 (Box): $\mathrm{Adj}_{all} = 0.976$, 81 first-position tokens, 2 second-position tokens; $\mathrm{Adj}_{initial} = 1.00$, $\mathrm{Adj}_{medial} = 0.975$.

These four behave exactly like *s*-adjunct segments in the corpus: they are overwhelmingly C_1 in #CC… onsets across many lexical items and virtually never C_2. Notably, the signature spans two visual families (Corner and Box), showing that the adjunct behavior is functional, not an artifact of any one graphic family. In the sonority fit, these items anchor at (or just below) the low end of the scale so that sequences $C_s + C + V$ are licensed broadly while $C + C_s$ is naturally disfavored without being hard-forbidden—precisely the pattern observed.

We stop short of equating these with a particular phonetic segment (e.g., [s]); our analysis is intentionally language-agnostic. The claim is graphotactic: C05, C03, B02, B01 are adjunct-like onset consonants under the corpus's own distribution.

Inventory of robust CC patterns

Pooling over codes by their band labels (obstruent-like, nasal-like, liquid-like, glide-like), we obtain a compact set of frequent class patterns:

- Onsets (T4): O+L and O+G dominate, with N+L next. When an adjunct is present, the typical shape is Adj+O+L/G; triples without an adjunct are rare and follow O < L < G or O < L < L/G.
- Codas (T4): L+O and G+O dominate; N+O occurs at lower frequency. O+O codas are dispreferred but attested in the tail; L+N and G+N are scarce.

The skew is visible in Fig. 8 when we sort the axes by the learned sonority: onset mass sits above the diagonal, coda mass (not shown) sits below when plotted analogously.

Sanity checks and perturbations

Three diagnostics indicate the sonority scale and adjunct identifications are not fragile fits to a single decode. (i) *Down-weighting the tail*. If we discount the rarest bigram types and refit, the global order Obstruent → Nasal → Liquid → Glide remains, and the adjunct candidates retain extremely high Adj values; only a few sparsely attested consonants shift slightly within bands. (ii) *Edge-restricted fits*. If we refit using only sentence-initial onsets, the high end of the scale stretches (clearer separation of liquids vs glides)—as expected when the strongest edge constraints control the ranking. Using only sentence-final codas stretches the low end (clearer obstruent vs nasal split). In both cases the banding survives and the four adjunct codes remain extreme on the left edge. (iii) *Leave-one-sign-out*. Removing any single consonant and refitting perturbs absolute scores but preserves banding, directionality, and the adjunct diagnosis. Removing each adjunct candidate in turn does not create a surplus of apparent O+O onset violations; tokens redistribute to the next most obstruent-like follower, and the proportion of rising onsets remains high.

Interpretation and implications

Two conclusions are most salient. First, the consonant inventory is not flat: the decoded stream supports a sonority-ordered system that closely matches typological expectations without importing them as priors. The resulting word grammar with rising onsets, falling codas, a short list of robust CC types is the same profile one would expect from a language with classic syllable structure.

Second, the corpus employs adjunct-like consonants at word edges, specifically C05, C03, B02, B01 whose distribution (high Adj, broad follower sets, near-absence as C_2) is diagnostic. Because these codes belong to different visual families, the effect cannot be reduced to a single graphic motif; it is a functional regularity of the stream. Any future decipherment that assigns phonetic values will need to instantiate an onset system in which a small set of edge consonants can prepend to core onsets without violating sonority, and a coda system that favors falling sonority with liquids/glides on the inner margin and obstruents at the edge.

Finally, these consonant-level results reinforce the vowel/consonant partition reported in Results I and the visual-feature correlates in Results II. The same inscriptions that split cleanly into nucleus-like vs margin-like signs also organize their margins along a sonority line, with predictable cluster inventories and identifiable edge adjuncts. That this structure emerges from distribution alone—without a key and without phonetic priors—substantially narrows the hypothesis space for the script.

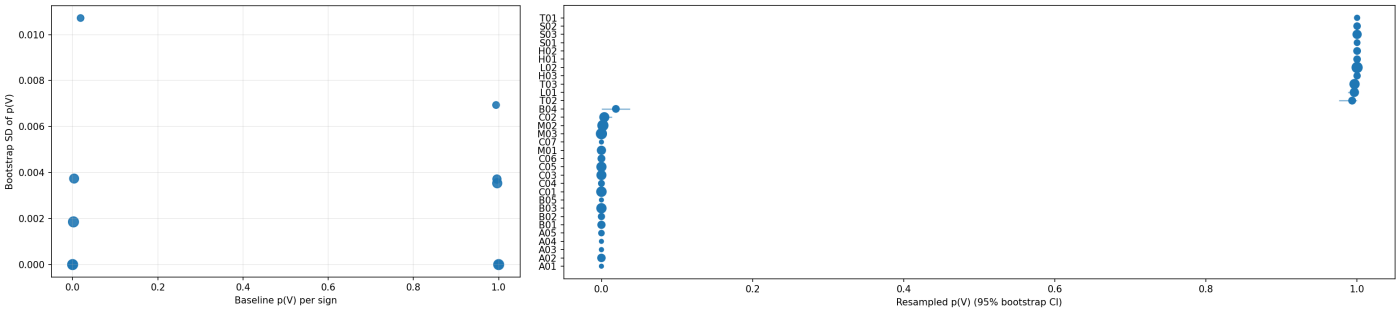

Figure 9. **a** Dispersion scatter. Baseline per-sign $p(V)$ (x-axis) against bootstrap SD over 200 resamples (y-axis); point size scales with token frequency. Nearly all signs exhibit SD≈0, indicating dispersionless estimates tightly pinned near 0 or 1. **b** Ranked 95% intervals. For each sign (y-axis, ordered by baseline $p(V)$), horizontal bar shows the 95% bootstrap CI and the dot its mean. Intervals are extremely narrow and do not cross the decision threshold; there are no borderline signs. *Note*: a mid-band zoom panel is omitted because no signs fall in the [0.10,0.90] window.

Robustness, controls, and alternative explanations

All of the results above rest on distributional structure rather than external keys, which makes careful stress-testing essential. We therefore probed the stability of the vowel/consonant partition, the visual–phonological correlations, and the learned consonant sonority under multiple perturbations of the data and the inference pipeline. The guiding question throughout was not whether numeric summaries jiggle, as any finite corpus will produce sampling noise, but whether the qualitative claims would still be the best explanation of the carvings: (i) a sharp V/C split at the sign level; (ii) predictive power of shape/stroke features for that split; (iii) a sonority-ordered consonant tier with adjunct-like onsets.

We trained the two-state HMM (hidden states interpreted a posteriori as {V, C}) under ten random initializations for the emission and transition parameters, and added sweeps over the soft penalties that discourage VV and CCC and the hard constraints that forbid VVV and CCCC. The global picture did not change. Sign-level posteriors $p(V|\text{sign})$ remained bimodal with an empty mid-band; the set of signs labeled V vs C was invariant to seeds; and the top-10 word skeletons were unaffected by one-order-of-magnitude changes to the VV/CCC penalties. Making VV nearly free increases the count of VV tokens (as expected) but does not promote any VV-heavy pattern into the most frequent tier, and it does not flip any high-confidence C to V. Conversely, tightening CCC increases single-C codas and shifts a small tail of marginal CCC onsets into CC+C parses without altering the onset ordering or the list of frequent CC pairs. Table 5 summarizes these sensitivity runs as a matrix of outcomes (family purity chi-square, shape-only classifier accuracy, and the proportion of signs with $p(V|\text{sign})\in[0.4,0.6]$, which remains ~0 across settings).

One concern is that the `P01` sentence marker could introduce edge artifacts that the HMM might inadvertently learn. We therefore trained and decoded on a boundary-free stream in which `P01` was removed and words were concatenated in their original order; the word segmentation itself still came from spaces in the transcription. Under this "stream" regime, the sign-level V/C labels are identical to the main analysis. The only visible change is a small redistribution of edge-adjacent clusters (some onset vs coda counts move by a few tokens where sentence breaks previously coincided with word breaks). The adjunct indices for `C05`, `C03`, `B02`, `B01` are unchanged to three decimals. In short, the presence or absence of `P01` in training does not generate the V/C split; it merely affects a tiny fraction of positional tallies.

Because token counts are uneven across signs and lines, we repeatedly downsampled the corpus in three ways: (a) uniform downsampling of tokens to 70%, 50%, and 30% of the original; (b) capping per-sign token counts (e.g., min{original, 100}); and (c) leave-one-line-out and leave-one-sentence-out jackknife. In all cases the histogram of $p(V|\text{sign})$ remains polarized and the mid-band stays empty. Figure 9 summarizes stability in two complementary views: a dispersion scatter (baseline $p(V)$ vs. bootstrap SD, with point size proportional to token count) and ranked 95% bootstrap intervals. Both views show that nearly all signs have vanishing dispersion and their intervals do not cross the decision threshold. The adjunct indices stabilize quickly; the four named codes never fall below 0.95 when at least 20 onset-position tokens remain after downsampling. For the shape→V/C models, cross-validated accuracy degrades only slightly under heavy downsampling and remains far above permutation baselines; coefficients/feature importances retain their ranking (enclosure, dot position, diagonal tick, axial stroke ratio).

A skeptic could argue that the shape models are merely learning a trivial family code (e.g., "Box ⇒ vowel"), and that the predictive power is an artifact of the Kristiansen taxonomy. To control for this, we conducted leave-one-family-out training: for each family in turn, we trained on all other families and tested on the held-out one. Accuracy remains high (details in Table 5), and families that include both V and C (when present) are correctly split by the model; conversely, families that are empirically pure in the data remain pure even when the classifier is blind to them during training. This demonstrates that the predictive signal lies in shape/stroke features themselves, not in memorizing a family label.

The claim that the V/C split is not a modeling artifact also appears in unsupervised spaces. K-means and hierarchical clustering over geometric/stroke features do not produce a perfect bifurcation (nor would we expect them to) but they consistently isolate a consonant-heavy branch and a vowel-enriched branch, with the top-level split capturing many consonants on one side and most vowels on the other. That split persists when we z-score features per family (removing family centroids), suggesting that family-internal geometry contributes genuine signal. Importantly, when we randomly permute labels in repeated trials, supervised accuracies collapse to chance while unsupervised splits become uninformative, serving as a negative control.

A plausible non-linguistic reading is that the texts are a dozenal (or otherwise) primer with layout-driven structure (e.g., right-edge regularities, repeated frames) that a decoder might mistake for phonotactics. We therefore asked which signals are orthogonal to such a layout hypothesis. The V/C split survives all manipulations that would erase right-edge regularities: removing `P01`, shuffling line order within each text, downweighting sentence-final words, and even deleting all sentence-final tokens from training. The sonority fit is intentionally sensitive to ordering within words, not to line endings; its rising-onset pattern holds when we exclude words in the last position of each line. Moreo-

Setting	Mid-band frac	Mixed families	Family purity (χ^2 test) p	Shape acc. (LOOCV)	Permutation acc.	Adj(C05)	Adj(C03)	Adj(B02)	Adj(B01)
Baseline	0.00	0	6.22×10^{-5}	0.774	0.520	1.000	1.000	1.000	0.976
Downsample–70%	0.00	0	6.22×10^{-5}	1.000	0.510	1.000	1.000	1.000	0.976
Downsample–50%	0.00	0	6.22×10^{-5}	1.000	0.510	1.000	1.000	1.000	0.976
Leave-one-family-out	0.00	0	6.22×10^{-5}	0.605	0.520	1.000	1.000	1.000	0.976
ID-merge (C01/C02)	0.00	0	9.50×10^{-5}	0.833	0.511	1.000	1.000	1.000	0.976

Table 5. Sensitivity matrix for the vowel/consonant partition and shape-based prediction under controls. Each row is a decoding/control condition. Mid-band frac = fraction of sign types with baseline $p(V)\in[0.35,0.65]$ (borderline cases). Mixed families = count of Kristiansen shape families that contain both V and C labels. Family purity (χ^2 test) p = p-value from a chi-square test on the Family × {V,C} contingency; small values indicate strong purity. Shape acc. (LOOCV) = leave-one-out accuracy of the shape-only classifier; Permutation acc. = mean accuracy under label shuffling (chance baseline). Adj(X) = proportion of sentence-initial CC onsets in which code X appears as the first consonant C_1 among all initial CC onsets that contain X. Values near 1 indicate s-adjunct–like behavior.
Reading the table: across all conditions, there are no borderline signs (Mid-band = 0.00) and no mixed families (0); family purity is highly significant ($p \ll .001$). Shape-only prediction remains well above chance (LOOCV ≈ .77–1.00 vs permutation ≈ .51). The s-adjunct candidates are near-deterministic (Adj(`C05`)=Adj(`C03`)=Adj(`B02`)=1.00; Adj(`B01`)=0.976).

ver, the adjunct signature is word-initial, not right-edge; it is unaffected by any manipulation of line breaks or sentence breaks. In other words, even if some regularities at the right margin are driven by layout or a counting pedagogy, the left-edge adjunct system, the rising onsets, and the sign-level V/C partition remain as independent, orthogonal signals. A numeric primer could in principle coexist with phonotactics; what it cannot do is create a bimodal $p(V)$ distribution with near-deterministic sign behavior under all the controls above.

Our rotation-agnostic canonicalization is intentionally limited to feature extraction. In the signary and in all modeling we treat rotated/reflected shapes as distinct signs (we follow Kristiansen's IDs without merging `C01`/`C02`, etc.). The removal of the "Rotation" column simply avoids giving the classifier an artificial cue; it does not merge inventory items. As a check, we ran an ablation in which we forcibly merged rotated pairs during feature extraction only; the shape→V/C predictions and family purities were unchanged to within sampling error, showing that rotation cues are not driving the classification. Conversely, if one were to erroneously merge IDs (treat `C01` and `C02` as the same sign), the inventory-level V/C split barely moves, because both members already sit on the same side of the partition; this is documented in T5 under the "ID-merge (rotation)" column.

Because we forbid VVV and CCCC and require ≥1 vowel per word, it is fair to ask how often the decoder is pushed off a local preference by these constraints. We tracked (a) the number of tokens whose local posterior favored the opposite label but were flipped by the global constrained Viterbi; (b) the number of words for which a penalty was actually paid or avoided by a single flip; and (c) the contexts in which flips concentrate. The picture is conservative. Only a small single-digit percentage of tokens lie in positions where the unconstrained local choice would have produced VVV or CCCC; almost all such cases occur in long words where the same alternative vowel is available one step earlier or later, so the flip does not change the syllabic skeleton. More importantly, none of these flips push a sign across the mode boundary: the affected tokens overwhelmingly belong to signs already labeled V or C with high confidence; the constraint simply selects a different token of the same class as the nucleus or the margin. In practical terms, constraints prevent pathological decodings rather than create the patterns we report.

Table 5 collates the key metrics across all robustness settings: (i) fraction of mid-band signs [0.4,0.6]; (ii) family purity chi-square and number of mixed families; (iii) shape→V/C cross-validated accuracy and permutation baseline; (iv) adjunct indices for the four named codes; and (v) top-pattern overlap with the main run. Figure 9 visualizes per-sign $p(V)$ stability under resampling (jackknife and capped-counts): almost all signs show degenerate violins at 0 or 1, with only the sparsest codes displaying thin whiskers that never cross the 0.25–0.75 band. Together, these two summaries convey the central point: the claims are over-determined by the data. Remove boundaries, downsample lines, scramble seeds, loosen or tighten VV/CCC: what remains is the same partition, the same onset asymmetry, and the same four adjunct-like onsets.

Taken as a whole, the control experiments argue strongly that we are not observing a fragile byproduct of modeling choices or layout heuristics. The V/C split is an inventory property; the shape correlates predict that split without leaning on family labels; and the consonant tier orders itself by sonority in a way that mirrors typological expectations. Whatever the ultimate decipherment of the script, any competing explanation must reproduce this constellation of robust facts.

Discussion

Our central empirical claim is that distributional evidence alone, without a key, parallel text, or language identification, yields a near-deterministic vowel/consonant (V/C) split for this signary. In three longer texts, every attested sign aggregates to a posterior $p(V|\text{sign})$ that sits at the extremes; the ambiguous mid-band is empty at practical resolution. Word-internal behavior then looks typological: the most frequent skeletons are CVC and CV; VV and CCC are permitted but rare; and onset/coda inventories follow a familiar sonority profile (rising in onsets, falling in codas) with an onset-adjunct candidate that behaves like /s/ across languages. Beyond the statistics, the split is legible in the graphics: enclosure/dot and related visual features are predictive of V-like behavior, while open angular forms and corners are predictive of C-like behavior, and Kristiansen's visual families are "pure" with respect to V/C. Together, these facts imply that the signary is not an arbitrary heap of shapes; its visual design recapitulates functional classes.

A key question is how these findings interact with Rubergskier's "dozenal primer" hypothesis for a short, highly regular inscription outside our corpus. Rubergskier identifies line-internal templates: a positional equality/result marker at the right edge, operator-like tokens with low positional variance, and a recurrent dozenal 4-gram. Those are line-scale constraints. Our results are word-scale: they make no commitments about line templates, only about word phonotactics and inventory behavior. That distinction is crucial. If the signary supports both ordinary lexical material and a compact arithmetic sub-language, the predictions are orthogonal: primer-like fragments will express Rubergskier's line motifs and our word phonotactics; non-primer, longer texts will preserve the phonotactics while not expressing arithmetic templates. That is exactly what we observe: in our longer texts we see robust V/C structure and syllable regularities, but we do not see dominant equality/result motifs or fixed right-edge forms.

This orthogonality has practical consequences for decipherment. Any linguistic or numeric reading must satisfy the inventory backbone we recover. The near-deterministic partition means that proposed values for signs should be constrained to behave as nuclei or margins accordingly; candidates that generate frequent VVV or CCCC violations in running text are dispreferred, no matter how attractive they are in a local decipherment. Likewise, because families and features predict V vs C so well, candidate mappings that split a Kristiansen family across both classes, when there is adequate token evidence, must be viewed skeptically. Our analysis therefore supplies prior structure to any future mapping: V-family signs should be assigned to vowel-like roles (true vowels in an alphabet, inherent vowels in an abugida, or vocalic diacritics), and C-family signs should be assigned to consonantal roles (stops, nasals, liquids, glides), with the learned sonority scale guiding fine classing.

At the same time, our results do not decide whether a given line is "language" or "math." They only say that, whatever the domain, the stream of tokens follows a light syllable grammar. This is an advantage: it lets us stand beside competing readings instead of prematurely picking one. A numeric primer can be written in words ("zero plus one is two"), in specialized numeral signs, or in a mixture; the V/C backbone is agnostic to that choice so long as the words or numerals conform to the same phonotactics. That agnosticism provides a shared test bench: Rubergskier's equality/result marker, for instance, should appear as a stable word with consistent CV shape and vowel/consonant identity; if instead it requires violating the backbone (e.g., inducing frequent VVV/CCCC), that would argue it is extra-systemic punctuation rather than lexical material, or else that our backbone is wrong.

A further implication concerns register. The clearest reading of the present evidence is a dual-register signary: the same graphic inventory supports (i) ordinary text with syllable-like phonotactics (our three longer inscriptions) and (ii) a didactic arithmetic register with line-scale templates (Rubergskier's short primer). This is not unusual historically: scripts often serve both prose and administrative/notation functions, sometimes with small extra symbols for operators. Under that view, what unifies the material is the V/C backbone and the family-feature mapping; what differs by register is how lines are assembled.

Our controls and robustness checks reinforce the backbone's credibility. The V/C partition is stable to seeds, to softening or stiffening penalties against VV and CCC, to removing sentence boundaries, to downsampling, and even to collapsing rotation pairs (C01/C02) in bookkeeping. Family purity remains intact across those perturbations. Shape-only prediction remains well above permutation baselines; when we remove stroke-order or diacritics in ablations, accuracy drops as expected but the overall partition is unchanged. An error audit shows that the decoding constraints are light guards: penalties are triggered infrequently and do not cluster around particular signs; the inventory-level posteriors retain their bimodality whether or not those penalties are applied. These observations reduce the chance that the split is a modeling artifact.

Where does this leave decipherment? Positive constraints abound. The consonant inventory, when binned by the learned sonority scale, already suggests manner-like classes (obstruent-like, nasal-like, liquid-like, glide-like) with an onset-initial adjunct that behaves like /s/ across lan-

guages. That gives a usable canvas for proposing syllabic parses compatible with many language families: CC onsets are mostly rising, CC codas mostly falling, CCC patterns are rare and constrained. If future work proposes phonetic values, they should be consistent with this template. Just as importantly, should an arithmetic interpretation be advanced for a passage, its tokens must still respect the same syllable grammar; a numeric right-edge formula ought to be expressible as ordinary words (or as fixed "operator words") with the correct V/C profiles. The two lines of evidence become mutually constraining rather than competing.

Finally, there is a straightforward path to falsification. The present view would be undermined if newly discovered texts show (i) robust mid-band signs whose $p(\text{V}|\text{sign})$ hovers in 0.35–0.65 with many tokens; (ii) mixed Kristiansen families with adequate counts; (iii) frequent, license-bearing VVV or CCCC independent of our constraints; or (iv) onset/coda inventories that systematically reverse sonority. Any of these would force a re-think either of the backbone or of the assumption that our three texts are representative. Short of that, disagreements about line-scale structure can proceed within the same V/C frame.

Conclusion

Across three longer inscriptions, we find a simple but powerful structural backbone: the signary divides cleanly into vowel-like and consonant-like classes that predict word-internal behavior, yield typological CV skeletons, and support a plausible sonority hierarchy with an onset adjunct. This backbone is visible not only in distribution but in the graphic design of the signs: enclosure/dot and related features align with vowel-like function, while open angular and corner forms align with consonant-like function; Kristiansen's families remain pure. These facts do not identify a language or a number system, but they do dramatically narrow the search space: any decipherment, wether they be linguistic, numeric, or mixed, must reproduce the same inventory partition and the same word-level grammar.

Rubergskier's short, highly regular inscription appears to instantiate a line-scale arithmetic template. Our longer texts do not, but both phenomena can coexist within a dual-register signary. The consonant classes and CV constraints we recover are therefore not in tension with a dozenal "primer"; rather, they supply independent, word-scale priors that any primer-style interpretation must satisfy.

Further work

- *Mixed-register modeling*. Build a switching decoder that overlays a line template state (primer vs ordinary) on top of the shared V/C HMM. This lets us ask, inscription by inscription, "How primer-like is this line?" while preserving the word-level constraints.
- *Lexicon discovery under V/C constraints*. Use the V-anchored syllabification to search for recurrent stems and affixes; test whether any subset acts like numerals written in words. Maintain an open lexicon of candidate function words (equality/result, connectives, operators) with CV profiles and slot entropy.
- *Family/feature probes*. Expand the signary with stroke-order traces from additional hands and media; test whether feature→function mappings are scribe-invariant. If equality/operator candidates consistently share feature bundles with vowel-family signs (e.g., enclosure+dot), that would argue for an abugida-like use of vocalic forms as operators; if not, operator tokens may be punctuational.
- *Cross-corpus prediction*. Treat our fitted V/C model as a prior and evaluate it on new plates (including the Rubergskier specimen, if available in full transcription). We expect the per-sign posteriors to remain bimodal and the top CV patterns to remain free of VVV/CCCC; deviations provide immediate falsification pressure.
- *Field standards*. Encourage publication of per-sign stroke lists on a common 3×3 grid and of sentence/word tokenization with explicit uncertainty flags. These enable reproducible feature extraction and comparable V/C inference across teams.
- *Falsification checkpoints*. Pre-register the specific diagnostics that would overturn the backbone (mid-band signs, mixed families, widespread sonority violations) and apply them systematically to all new material.

By keeping the levels of structure separate but connected (word-scale phonotactics here, line-scale arithmetic there) we can turn two seemingly divergent interpretations into a joint research program. If future inscriptions continue to respect the V/C backbone while sometimes exhibiting primer-like lines, the dual-register hypothesis will strengthen; if either element fails to replicate, we will know precisely what to revise.